Kristiina Nieminen

AN IMPROVED INFORMATION MODEL FOR KNOWLEDGE-BASED WORK

How data is treated in process and what is quality of results?

Helsinki Metropolia University of Applied Sciences

Master's Degree

Master's Degree Programme in Business Informatics Thesis

31st May 2018

Helsinki Metropolia University of Applied Sciences

Author(s) Title Number of Pages Date	Kristiina Nieminen An improved information model for knowledge-based work 82 pages + 5 appendices 31st May 2018
Degree	Master's Degree
Degree Programme	Master's Degree Programme in Business Informatics
Specialisation option	*Name of the specialisation option*
Instructor(s)	Heikki Rouhuvirta, Senior Adviser (retired), Statistics Finland James Collins, Senior Lecturer of Organisational Development & Management Research, Metropolia University of Applied Sciences

Since the 18th century, statistics of the Finnish society have been compiled in Scandinavia. In the beginning, compiled statistics covered information about the population but over the decades new subject areas such as economy, living conditions, environment and business sectors have been included into statistics production.

Statistics are compiled to support decision-making, and therefore they must meet the set quality criteria. In order to ensure this obligation, statistical producers need descriptive information, the so-called metadata, of statistical data and the compilation process. The aim of defining and using metadata is to enable the users and statisticians to identify and to understand the key qualitative factors related to data and results.

The purpose of this Master's thesis was to create an information model that may be used to store detailed information about the process and the data. The first task was to analyze the current practices and information needs, the production guidelines of general statistics and the available information models. Then the analysis results were used in the design of an improved information model. The new information model complements the existing common metadata system in Statistics Finland, and creates a completely new information content for statistical process related information.

The research method was an inductive, data-based analysis that was based on a survey questionnaire about the current status and information needs. The respondents were randomly selected from Statistics Finland's statistical units. Based on the responses, a list of

the current production system's weaknesses was created. In addition to this, a literature review covering international development work, research reports and generic information models was compiled. The theoretical framework of this thesis was founded on ethical principles, legislation, working instructions composed by international actors and EU quality requirements.

An information model design was accomplished from the point of view of process related information. The designed final information model is a process-based model, i.e. it brings together the key elements of statistics production: the process and its phases and steps, statistical data, statistical methods, rules and actors.

As a results the designed information model is simple and general in nature, so it may be applied in other national statistical institutions as well. The structure of the final information model is hierarchical, whereby dependencies between the different elements are identified, described and clearly understood.

Keywords	Information model, gsbpm, cossi, gsim, process, statistical data, metadata, statistics

Contents

Definitions and Key Terms

- “statistics” means quantitative, aggregated and representative information characterising a collective phenomenon in a considered population (European Union 2010),
- “production” means all the activities related to the collection, storage, processing, and analysis necessary for compiling statistics (European Union 2010),
- “data collection” is gathering data from all types of sources, such as statistical surveys, questionnaires, administrative records. Data is collected taking into account the quality, timeliness, costs and the burden on respondents (European Union 2010),
- “statistical unit” means the basic observation unit, namely a natural person, a household, an organisation and other undertakings, referred to by the data (European Union 2010),
- “data” is a collection of measured observations regarding the statistical unit. Data is organised as an observation matrix, where rows typically represent different repetitions of an experiment (observations), while columns represent different types of data (variables),
- “statistical process” is the process flow where collected data is processed according to the agreed methods. Processing goes through main phases: Design, Collect, Process, Analyse, Disseminate, Evaluate,
- “metadata” is data (information) that defines and describes other data. Statistical metadata are defined as data about statistical data, and comprise data and other documentation that describe objects in a formalised way .(SDMX 2016),
- “information” is data that is accurate and timely, specific and organized for a purpose, presented within a context that gives it meaning and relevance, and can lead to an increase in understanding and decrease in uncertainty. (Business Dictionary n.d),
- Matrix is a collection of sells (values) arranged into a fixed number of rows and columns. Statistical data is often presented in a form of matrix

1 Introduction

1.1 Overview

Official statistics have been produced for over 150 years in Finland and elsewhere even longer. During these years several, alternative methods have been used to collect data and to compile statistics. Therefore, there has been a continuous need to improve and to develop existing practises. This necessity became even more necessary, when personal computers were adopted to daily work in enterprises and national statistical institutions, such as Statistics Finland. Still, basic principles of statistics compilation are the same as before personal computers.

United Nations Statistics Divisions (United Nations Statistics Division 2014) has defined the meaning of official statistics as follows:

Principle 1. Official statistics are those providing

> an indispensable element in the information system of a democratic society, serving the Government, the economy and the public with data about the economic, demographic, social and environmental situation

So, the main role of official statistical institutions is to provide statistics that are produced professionally and independently, and meet the quality criterion set by UN Statistics Commission. In Finland, official statistics describe Finnish society comprehensively, so that released data is reliable, impartial and timely. (Official Statistics of Finland 2013)

To ensure comparability and quality of statistics the national statistical institutions and international actors have created global frameworks, practices and instructions. To name few of the common practices, composed in co-operation, are the System of National Accounts (SNA) and the Consumer Price Indices (Eurostat 2016a; Eurostat 2017; Eurostat 2013).

Some of these common practices are more general by nature as the developers have been international actors such as the United Nations Economic Commission for Europe (UNECE), the International Labour Organization (ILO) and the International Monetary Fund (IMF). More detail instructions and practices have been set by the European Union (EU) in co-operation with its members.

These jointly agreed practices cover instructions, manual, common methods for data processing and regulations. The aim in defining common practices is to offer comprehensive, detailed instructions in order to help statisticians to produce comparable and high-quality statistics.

In worst case, without any common practices and standards, every statistical department could make their own decisions how to process the data, causing unnecessary variance in the results. This may impair comparison between countries. Most important of these practices are those regarding data collection and processing for these are phases that has the strongest impact on results and comparability.

Therefore in 2007 UNECE founded the Generic Statistical Business Process Model, GSBPM in order to offer common terminology and to divide process into phases (UNECE 2013a, p. 3). The GSBPM was afterwards adapted in national statistical institutions. This model helps to describe what is done in process phases and what are the similarities and differences in statistics production systems.

UNECE also founded the Generic Statistical Information Metadata model, GSIM, in order to draft common metadata elements used in the GSBPM (UNECE 2013b). A challenge at the moment is to combine common practices, business process model and metadata models with daily statistical work and to ensure quality of results.

1.2 Business Challenge

Enthusiasm to examine salient statistical information and to develop and information model comes from the daily challenges. Daily practices has shown that there is too little standardised and structured information captured and stored from data processing. Therefore personal ambition is to create an ideal information model that provides standardised structure for all metadata regarding statistical data and process.

Idea is that an ideal information model combines together normative requirements, structured metadata, unified practices and the generic process model with real life statistical production. Ideal model will cover the main process phases where actual data is manipulated: Data collection, Processing, Analysing and Disseminating.

In the future with this developed information model statistician may manage compilation process as a whole. This helps quality managers, researchers, analysts and developing party external users to get sufficiently information about decisions and methods used in the released statistics.

Detailed information about statistical data, process and statistical methods is captured and stored following structure of an information model. This information may be retrieved and utilised later in process. Afterwards collected metadata may be used to summarise production rules and quality of results.

This kind of combined generic level information and process model has not yet been planned in Statistics Finland. So, the aim is at contributing to overcoming this deficiency.

1.3 Case Company Presentation

The research is carried out in the national statistical institution in Finland. Statistics Finland was established in 1856 to produce statistics at first primarily for government needs but later on also for private companies and citizens that use statistics in their work.

Statistics Finland operates under the Ministry of Finance but is independently responsible for its activities, services and produced statistics. Statistics Finland provides reliable information concerning social and economic conditions for decision-making, research and inhabitants at large. Statistics Finland compiles approximately 75% of Finnish official statistics from 26 different topics covering nearly 200 sets of statistics. In the bureau works approximately 795 employees. (Statistics Finland 2016a)

Statistics Finland approved the GSBPM in 2009 in order to enhance quality management principles and standardise terminology (Statistics Finland intranet 2016a). The aim is to use the model as a basis for describing statistics process flows, in working instructions, in designing storage system folder structure and in design of user-interfaces for statistical data processing in in-house tailored systems.

Implementation of the GSBPM has been going on already for few years. It has been a challenge to adapt new standards while using diverse and in some cases obsolete IT-systems that do not bend that easily to new working methods.

The figure 1 presents structure of the GSBPM-process model. The GSBPM model contains eight main phases of which the most important phases, from the statisticians' point of view, are Collect, Process, Analyse, Disseminate.

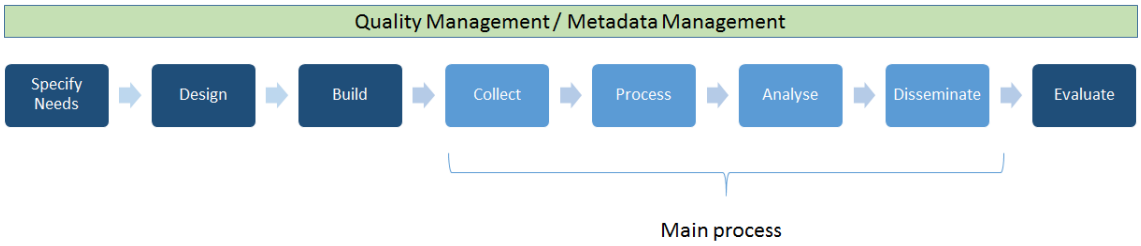


Figure 1. The overall Statistical Process Modell at Statistics Finland

In the figure 1, above the process phases is an element called "Quality and the metadata management" that is overarching, comprehensive, element. Quality and metadata issues need to be acknowledged in entire process flow irrespective of phase.

In this research the main concern is put on the phases where statistical data is processed. These phases are presented in the figure 1 with light blue colour.

1.4 Objective and Scope

Statistician need to understand thoroughly input data, data processing methods and quality of statistics. In order to achieve this demand, statistician need detailed information about definitions, concepts, variables, key indicators and statistical figures as well as background information about how data was collected. Also information about statistical methods including processing rules, parameters and formulas as well as order of processing steps is necessary.

Therefore, it is important to capture enough descriptive information about data content, process and quality indicators, and to utilise this information in next process steps. This way statistician may answer to question that matters to him most: *How is data treated in process and what is quality of results*.

Given this, the objective in this research is to design a new information model that may be used as a structure for statistical information. After this, modernised information model

may be implemented to statistical production, statistician may examine and investigate collected metadata and to compare these with international instructions, legislation and quality indicators.

This helps statistician, who is responsible of statistics production, to ensure quality and to understand content of results. Idea is, that every statistician should have good understanding of their statistical data processing, processed data and methods introduced in process.

So, the emphasis in this research is to combine information requirements and common business process model into improved information model. Output of this research is an expanded information model that meet set requirements.

1.5 Thesis Process

In this thesis I will study information that describes data and process. The aim is to point out information that is needed to help statistician understand released statistics quality. This information is then structured as in hierarchical format, so that relation between concepts is clear.

The thesis work is divided into seven stages. Figure 2 demonstrates these stages highlighting those where input data is collected for analysis. The output of a stage is shown in the figure with dark blue colour. Outputs are carried forward from stage to next one.

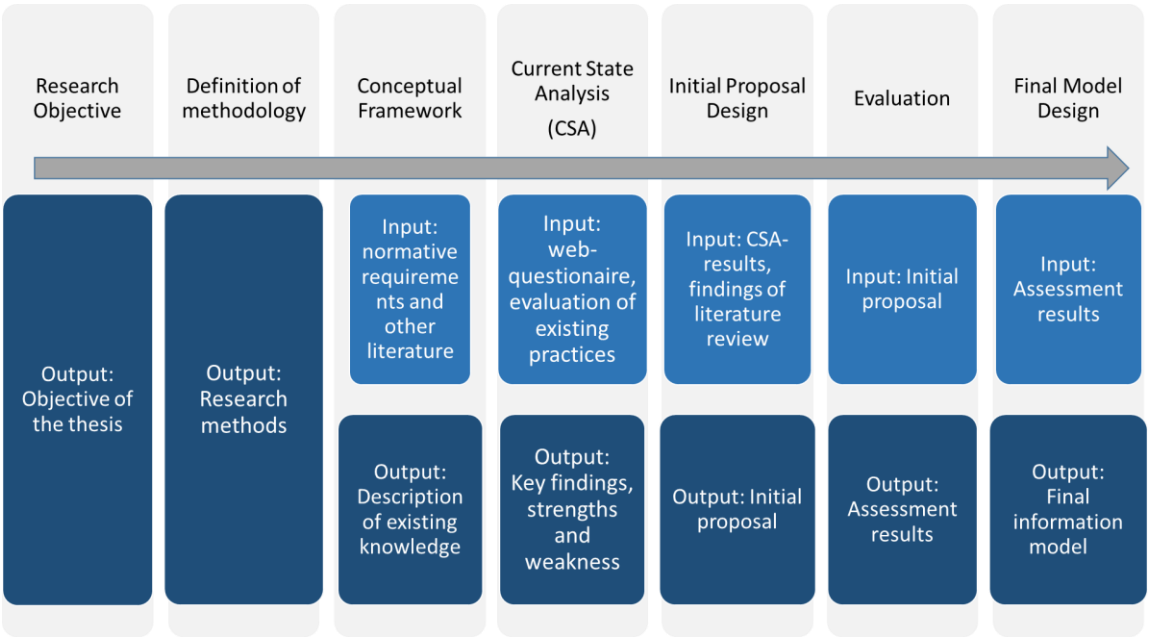


Figure 2. Outline of the thesis work-process

First stage, clarifies thesis objective and scope while second stage is needed to outline research method and to define research design that satisfies the aim of thesis.

Third stage, is needed for looking through normative requirements dedicated for statistic compilation and literature that describes business process- and information models. The aim is to outline conceptual framework that is later utilised in data analysis and model design.

In fourth stage, current state of metadata management in statistics is analysed and best practice already in use in Statistics Finland are examined. Analysis results is a list of requirements and identifiable weaknesses in current processes and information management.

Fifth stage, is to combine a list of requirements and weaknesses with the existing process- and information models and international practices. This collection establish starting point for an initial proposal building. The initial proposal is designed at this stage in order to be able to evaluate it by a focus-group.

Final stage, is to finalise the designed model based on the focus-group assessment results.

1.6 Thesis Chapters

The chapters in this thesis are divided according to the thesis process, figure 2. In the chapter 2. Methods and Material, research approach and design are described generally. This chapter also delineate the data collection used for current state analysis, the proposal design and the proposal assessment in focus group.

The chapter 3. Existing knowledge covers literature review and as a result of it, summarises existing knowledge at this area. The literature review concentrates especially in the Generic Business Process Model and the common information models used in international and national statistical institutions. In the chapter 4. Current State Analysis, data collection methods and the analysis results are elaborately described.

The chapters three and four help us to recognise current strengths and weaknesses as well as to identify what is already known and designed to be utilised at statistics compilation. The current state analysis results are combined with the existing standards, the GSBPM and the GSIM complemented with COSSI-model (Statistics Finland 2003). This information is then further analysed and an initial proposal is designed based on these findings.

The chapter 5 describes the designed initial proposal and gives an overview to an assessment that was done by a focus group. The focus group feedback is used to improve the suggested model and to finalise the model content.

Design of the final information model, a solution for information management, is outlined in the chapter 6 and thesis conclusions are presented in the chapter 7.

Next chapter gives an overview to a research approach and design. It will also give general view to research methods that will be used to carry out the thesis objective.

2 Methods and Material

This chapter offer ideas of the research approach and actions that were needed before a final information model is designed. Research philosophy was done in a spirit of realism for this approach suits best for action research where different kinds of research methods are used.

2.1 Research Approach and Strategy

Research approach may be deductive or inductive depending on selected research philosophy. Deductive approach is used when you map out appropriate theories for your research question while inductive approach is opposite to this. Inductive approach lets researcher to create own theory from the observations, arguments and reasoning.

Hence, inductive approach was used in this thesis to create improved information model. This method may also be called as bottom-up for the aim is to generalise observations into information. Recognised weaknesses, demands and requirements were used as a starting point in an information model design.

Current state analysis is a method that is used to observe internal working environment and to diagnose what are organization's strategic capabilities to succeed in its' task. Analysis items are strategic capabilities such as use of resources and competences, leadership, competitive situation and rivalry on organization's operations (Johnson & al. 2015, p.59, 65).

Several methods are available for analysis of an organisation. To name few of these, there are a risk analysis, a simulation of activities and processes, a cost-benefit analysis, a SWOT-analysis and a GAP analysis[1] (Public Recommendation JHS 171 2009, chapter 5.1). In this research the SWOT-analysis was selected as analysis method for it categorises analysis results into four group: weaknesses, strengths, opportunities and threats. The focus was pointed only at identified strengths and weaknesses.

[1] GAP analysis is a method to compare how well organization performs at the moment compared to its potential (Cambridge Business English Dictionary, n.d.)

Research strategy was a mixture of two methods, survey and case study. A survey was used to gather requirements by analysing current status of statistics production. Data was collected with a questionnaire. A case study was used as additional method to gather examples from real life. Three examples were selected for a case study-analysis.

Selected research strategy is performed as an action research that is practical way to examine target and to compose a list of actions that solve identified problem or may be seen as current weakness.

2.2 Action Research

The aim in this research was to create information model that bind together normative framework, existing knowledge and current state analysis-results.

Approach method was based on the action research which is commonly done in iterative cycles. Especially in a stage of final model design, several iteration cycles were needed. Below, in the figure 3, is an illustration of a common action research process.

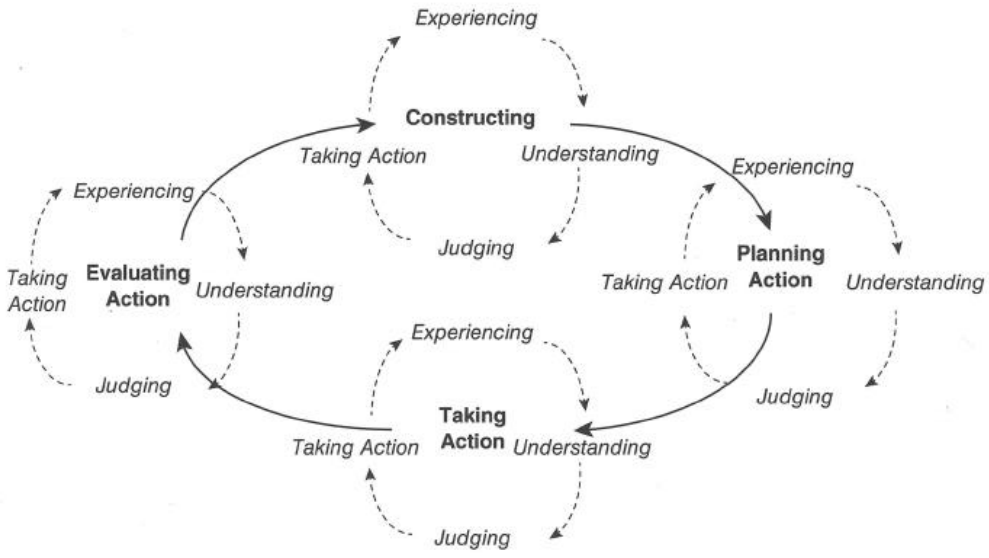


Figure 3. The General Empirical Method in Action Research Projects *(Source: Doing Action Research in your own organization, Coghlan & Brannic 2014, p. 30)*

The Action Research model contains four operations that has similar iteration steps, hence it was reasonable to divide this research design accordingly into following actions:

1. analysis of existing knowledge
2. current state analysis
3. design and assessment of an initial proposal
4. final information model design

Research process started with a literature review that was done to get a comprehensive understanding of the existing knowledge. The aim was to identify strengths in existing knowledge, but also recognize information gaps in presented information models. Literature review covered also normative documents that instructs statistician in their task to compose statistics. A list of qualification for information management and dissemination was deduced from normative documents. A conceptual framework is generated from the literature review results.

In the current state-analysis a web-questionnaire was primary method to examine current practices. Aim of the questionnaire was to gather information about

- how well statistician know the generic process model GSBPM and the statistical information model COSSI,
- how broadly statistician use these standardised practices in daily work,
- what are requirements that statistician set for detailed information

Supplementary information was then collected by examining the existing improved practices in Statistics Finland. Three case studies were selected: the Administered Data Collection, the Generic Editing Model and the Big-data processing. Beside these also Systematic Quality Audit-reports were reviewed to identify new ideas for information management.

The CSA results, the qualification list and the existing knowledge were combined together. Afterwards this summary list was utilised in design of an initial proposal.

Finally, an assessment meeting was organised with a focus group. In this meeting two methods, an assessment questionnaire and discussions, were used to collect expert

opinions of the presented initial proposal. The assessment results and received suggestions to improve the initial proposal were used in final step to complete the information model.

2.3 Data Collection

In this research, methods used in data collection was based on the decision made in research design. The research strategy outlines following data collection methods: arrange a survey, study three cases and analyse Systematic Quality Audit-reports for CSA. Second data collection cycle was arranged to gather expert opinions of the presented initial proposal.

The data one, that is combination of several collection methods, was collected for the current state analysis. The data two, an assessment results, was collected for finalisation of an information model design. The table 1 shows selected data collection methods and data sources.

Table 1. The data collection methods and sources

	Data collection method	**Document content**	**Data source**
Data 1a	Survey: Questionnaire	Questionnaire used for analysis of current practices of statistics process and metadata	Statistics Finland/ Statisticians
Data 1b	Case Study: Investigation of best practices and metadata elements	Case1. Administered data collection	Statistics Finland
Data 1c	Case Study: Investigation of the editing model descriptions	Case2. Generic Editing Model	Statistics Finland
Data 1d	Case Study: Evaluation of big-data processing	Case3. Big data processing -- Based on expert experiences	Kristiina Nieminen
Data 1e	Analysis of the auditing-reports	Systematic Quality Audit –reports	Statistics Finland intranet
Data 2	Survey: Assessment form and discussions	Assessment form used for the evaluation of the initial model	Statistics Finland/ Focus Group

Main data source was a web-questionnaire that was used to study current status. It covered 95% of collected information. Complementary information to outline the current status was gathered by studying documents concerning cases 1b and 1c, by studying an example of big data processing (data 1d) and by analysing the auditing reports.

Second data was collected with an assessment form. The focus-group members completed this form at the same time in the evaluation meeting.

Fuller presentation of the current state analysis, data collection-methods and analysis of the results are more closely opened up in chapter 4. Current State Analysis

2.4 Literature review

The aim in literature review is to understand the operational environment of statistics, to outline what is already known, existing knowledge and to bring forward set requirements in the field of this thesis.

Following paragraphs outlines topic that were selected for the literature review while the chapter 3 Existing knowledge presents the review results.

Normative requirements

There are several international and national principles outlining normative requirements for statistics production. In this research these ethical principles, legislation, requirements and recommendations were noticed and used as a framework of the research.

Components for the framework were collected from Statistics Finland's internal webpages (Statistics Finland 2016b). This list gives basic guidelines for daily work forming a research backbone and also laying benchmark for the current state analysis.

Literature on information management

Literature review was carried out to gather current knowledge about the Generic Statistical Business Process Model, GSBPM and the Generic Statistical Information Model, GSIM. Both models are internationally approved, providing standardised structure for information management.

Emphasis in this research was put especially to the information that need to be captured and stored from process and statistical data and additional information that may be later in the process retrieved and utilised. Target in this review was to highlight those practices that may be applicable in statistics compilation in Finland.

The aim, in this literature review was to outline strengths and weakness in existing knowledge: are there controversy areas of knowledge, missing knowledge or possibly imperfectly articulated concepts. The aim was also to recognize information components and field of knowledge that need further research.

2.5 Design and assessment of an initial proposal

The design process of an initial proposal started by combining solutions presented in the literature together with the results of the current state analysis. This list set requirements for an information model design.

The requirement list was used to design key concepts, sub-concepts and hierarchical structure for defined concepts. Key concepts [2]were defined for information covering process, data, quality, methods and for additional information that complements key concepts.

The idea was to draft an initial proposal model that combines process flow, actual data and information into a comprehensive solution giving enough detailed information for statistician to observe and to evaluate the process and its' results. The designed initial proposal was supposed to offer information about

- statistical data itself and of the changes made to that data,
- important indicators describing quality of data, statistics and process
- performance of the process

[2] In this thesis, from now on, term element is used in the same meaning as concept. An element here refers to the hierarchical structure that is needed to show the subordination of defined concepts.

Finally, the focus group had the chance to assess the drafted model. The aim was to verify if this initial proposal follow the international requirements and organization practices, and to evaluate if the initial proposal was implementable in the existing metadata systems.

3 Existing Knowledge

3.1 Overview

In the beginning of this research, it was necessary to carry out a literature review to highlight the essence of existing knowledge and to answer to the questions:

- *What are the international requirements composed for statistics compilation and especially for information management?*
- *How well international organisations have instructed national statistical institutions to capture and to utilise information, a.k.a metadata, in their statistics production?*
- *Are there suitable information models available that meet the set requirements?*

The aim is to carry out the literature review by investigating international and national instructions, information models and development work. The literature review begins with an orientation to normative requirements and formation of the research framework, and continue with the analysis of specific literature.

3.2 Normative requirements for Statistical Data Processing

Several international and national principles set requirements for statistics compilation, statistical data processing and for dissemination of information of statistical methods, rules and formulas. In this research ethical principles, legislation, requirements and recommendations were noticed in order to compose the framework of this thesis.

Documents for forming the framework were gathered from Statistics Finland's internal web-pages (Statistics Finland 2016b). The list, available in the web-page, provide basic guidelines for daily work and form the backbone in this research, also offering evaluation criteria for current state analysis.

Next we take a look at these documents, principles, legislation and practices. International authorities and actors such as the United Nations Economic Commission for Europe UNECE, the Organisation for Economic Cooperation and Development OECD and the European Union EU have composed these documents.

The Ethical Principles are composed by two authorities, the International Statistical Institute (ISI) and UNECE; the legislation is enforced by the EU Commission and the national authorities; the requirements and the recommendations are compiled by international authorities such as OECD, UNECE, Eurostat and Statistics Finland.

Main actors, from Statistics Finlands' point of view, are the EU Commission, the EU Council and the EU Parliament who has power to propose, adapt and enforce EU legislation (European Union n.d.) that need to be applied in Finland. All these authorities have, during the decades, compiled several instructions, manuals and recommendations for the statistics compilation yet in this research interest is put mainly on those documents that are referenced at SF's web-pages. Investigated documents are listed in the table 2.

Table 2. The normative requirements set for compilation of official statistics

Document type	Document content	Authority	Document reference
Principles	Fundamental Principles[3] (1994)	United Nations Statistics Division	UN Statistics Division 2014
	Professional Ethics[4]	International Statistical Institution	ISI 2009
Legislation	EU Legal Framework [5], Regulation 223/2009	Eurostat	European Union 2010; European Parliament and Council 2009
	Statistics Act 280/2004 [6]	Statistics Finland	Statistics Finland 2004
Practices	Code of Practice	Eurostat	Eurostat 2016a

These documents were investigated in order to highlight current requirements for statistical information. The aim in this analysis was to produce a qualification list that is useful when exploring questionnaire responses of the current state analysis.

The pre-analysis of the selected documents shows that the content of these documents is mainly too generic so no detailed requirements may be deduced from these. Though,

[3] The principle 3 sets requirements for information in the following way:
To facilitate a correct interpretation of the data the statistical agencies are to present information according to scientific standards on the sources, methods and procedures of the statistics.

[4] Topics of process and metadata are covered in the chapter 7 Exhibiting Professional Competence that sets requirements with words:
statistician shall seek to upgrade their professional knowledge.
In the chapter 9 Exposing and Reviewing Methods and Findings requirements are set with words:
statistician should provide adequate information to colleagues to permit their methods, procedures, techniques and findings to be assessed independently.

[5] Regulation 223/2009 states that
European statistics shall be developed in conformity with the statistical principles set out in Article 338 of the Treaty on the functioning of the EU and further elaborated in the European Statistics Code of Practice, namely, that: 'the production of Union statistics shall conform to impartiality, reliability, objectivity, scientific independence, cost-effectiveness and statistical confidentiality; it shall not entail excessive burdens on economic operators'.

[6] According to Statistics Act (2013), the objective it is
to ensure the availability of reliable statistical information required in social decision-making and planning and in fulfilling obligations under international statistical co-operation by harmonising and rationalising the principles and procedures applied in the collection, processing, use, release and storing of data, to promote the observation of good statistical practice in the National Statistical Service and to ensure that the rights of those who provide data for statistical purposes or whom the data concern are upheld. The purpose of the Act is also to extend the use of the data collected for statistical purposes in scientific studies and statistical surveys on social conditions.
(https://tilastokeskus.fi/meta/lait/index_en.html)

some common qualification may be derived for the use of this research. The list of qualification is presented in the table 3 below.

Table 3. The qualifications for statistics compilation

Source of qualification	**Qualification**
UN Fundamental Principles of National Official Statistics	Need to offer information of the data sources, statistical methods and procedures
International Statistics Institution, ISI	Need to provide information of statistical methods, procedures, techniques and findings
Legal framework of European Statistics and European Statistics System	Need to provide information of methods, procedures and techniques, and need avoid excessive burden on economic operators
European Statistics Code of Practice (Cop)	Need to identify strengths and weaknesses of the processes and product quality (systematically and regularly)
	Need to review, monitor and revise existing data processing practices (regularly)
	Need to assess and validate source data, intermediate results and statistical outputs (regularly)
Statistics Act	Need to inform respondents about the procedures used in the production of the statistics
	Need to use uniform concepts, definitions and classifications

Almost all documents emphasize the need to offer sufficiently information of the statistical data and the data processing. These qualifications are in line with the aim of this thesis.

3.3 The literature on the information management

International and national literature was reviewed in order to create an overall vision to the existing knowledge in the field of the GSBPM and the management of statistical information. The international development work on the field of information management have been in progress for years so there is enough documentation available.

UNECE released the Strategic Vision in 2011 (UNECE 2011a) that emphasizes the necessity to modernise processes and products if a statistical organisation desire is to remain relevant and sustainable. It also reminds that increasing volumes of data challenge the traditional practices. The Strategic Vision notes that

> The production of statistics should be based on common and standardised processes, transforming raw data into statistical products according to generic and commonly accepted information concepts.

(UNECE 2011b, p. 3-4)

The aim of the Strategic Vision is to understand the basic elements, so called cornerstones that are needed to industrialise and to modernise statistics production. (UNECE 2012, p.4). The cornerstone elements, in this vision, are statistical concepts, information concepts, statistical methods and technology. The GSBPM and the GSIM-model relate to statistical concepts and information concepts respectively (UNECE 2012, p.4) thus these are complementary models for the production and the management of statistical information.

The cornerstone vision was improved further to so called “Grand Unification” that was introduced in 2012. New approach was needed to bring together three disciplines: the subject-matter experts (business), the methodologists and the information technologists that commonly take part in the production development.

An illustration below, in the figure 4, presents key elements of the Grand Unification that contains four cornerstone element as well as the additional information elements for Business and Generalised Statistical Production System. (UNECE 2012, p.11).

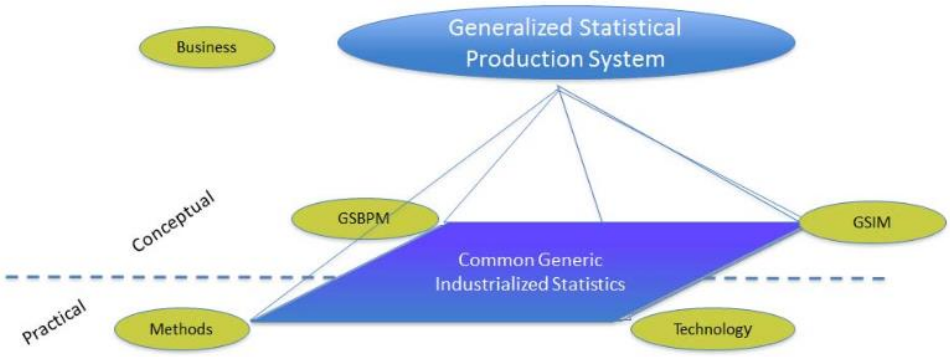


Figure 4. The Grand Unification (UNECE 2012, p.11)

The illustration above shows diversity of the key factors that need to be taken into account in the development of statistics production system. Practical factors are statistical methods and technology. Statistical methods are used when collecting and validating data, covering methods such as sampling, editing, imputing and estimating. Technology consists of IT-system components, such as databases, software, telecommunication applications, that are integrated as a unity that enables data processing. Conceptual factors are the GSBPM and the GSIM that may be used to describe processes and data, and to automate statistical data processing. The COSSI-model[7], that is developed in Finland, is an additional information model that is used for describing statistical data in Finland. Hence it ought to be investigated too in this context.

So, concentration in this research is put on the conceptual elements of the UNECE Strategic Vision and the COSSI-model. Main sources for the literature review are the material provided by UNECE, Eurostat[8] and Statistics Finland.

Beside the research questions presented in the chapter 3.1 Overview, strengths and weaknesses in existing documentation were explored. It was important to identify weaknesses in order to avoid making similar mistakes in the design of a proposal.

3.4 The GSBPM

References to the GSBPM has been made already several times in this thesis without more specifically explaining content of it. In this chapter the aim is to look closer at the process model and information management in it.

Important internationally qualifications set for information management and process are listed in the table 3. The qualifications for statistics compilation. According to the list, statistician need to provide information about procedures, data sources, introduced statistical methods and to identify weaknesses in the process.

Following sub-chapters present the GSBPM, the new indicators defined to be used with the GSBPM and the evaluation of the process model.

---

[7] Statistics Finland provides documentation on the COSSI-model in web-page https://www.stat.fi/org/tut/dthemes/drafts/index_en.html

[8] UNECE provides documentation on the GSBPM and the GSIM in web-portal https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0 and Eurostat provides information on metadata in web-page http://ec.europa.eu/eurostat/data/metadata

3.4.1 Overview

Purpose of the GSBPM is to describe and define the process flow where statistical data is manipulated and statistics are compiled (UNECE 2013c).

The GSBPM –model is based on the process model that was initially developed and used in Statistics New Zealand. In 2007 participants at UNECE workshop, entitled "Metadata and the Statistical cycle", agreed to take the Statistics New Zealand's process model as a starting point for the further model development. The process model development work continued so that in 2009 an improved generic business process model, the GSBPM version 4.0, was released. (UNECE 2013c, p.3)

According to UNECE report (2013c), the GSBPM is a reference model that is intended to be used by organization to a different degree. The model provides

> A standard framework and harmonized technology to help statistical organisations to modernize their production processes.

UNECE instructs statistician to use the GSBPM model as a reference framework that enable cross-disciplinary communication and understanding through common terminology (UNECE 2012, p. 11). By contrast, this framework does not offer detailed instructions how to apply the generic model in practice.

The figure 5 presents the GSBPM model that has eight main phase (blue colored boxes in the figure), the steps within a phase (red colored boxes) and the quality management layer (green colored box) on top of process phases.

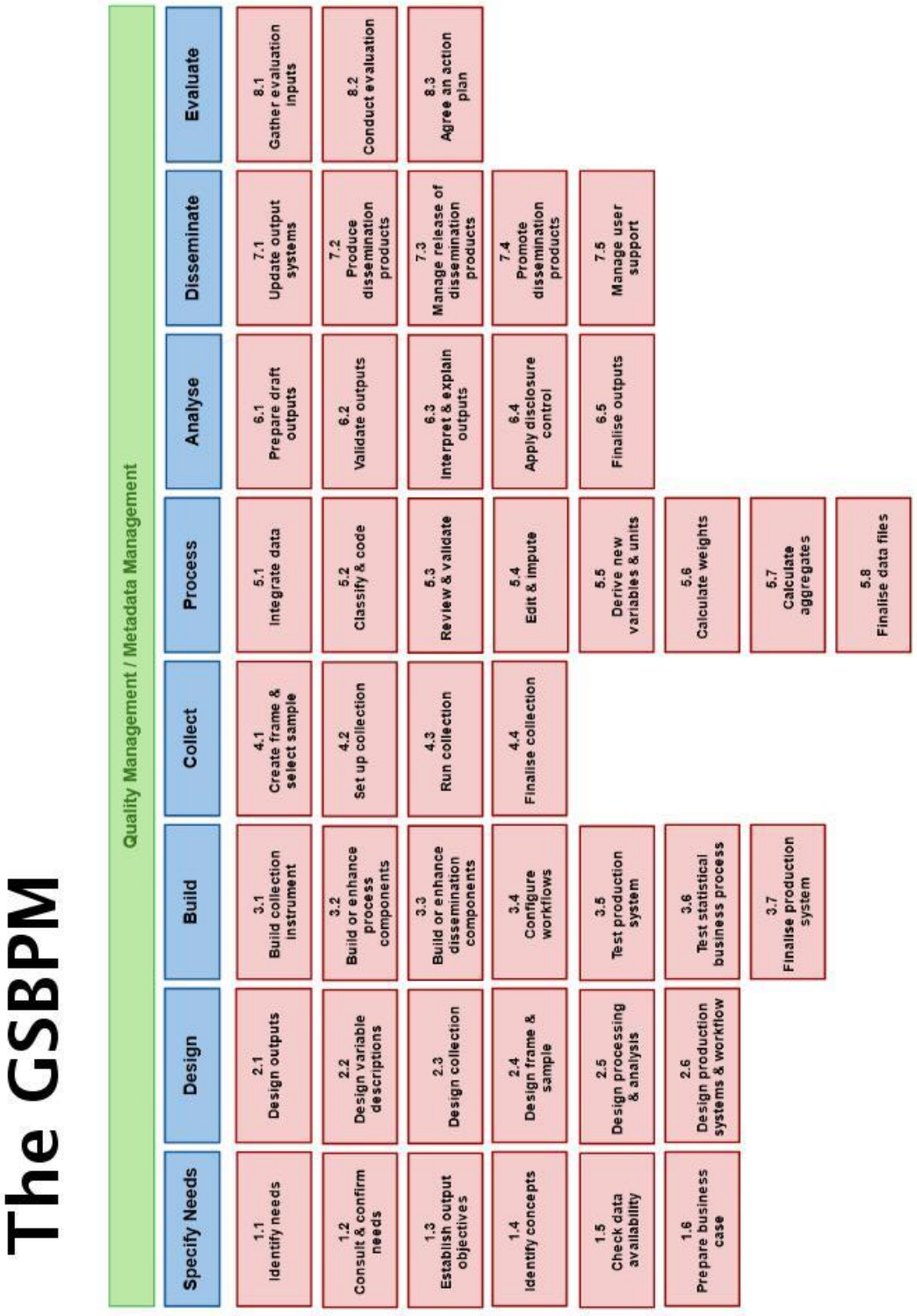


Figure 5. The Generic Business Process Model defined by UNECE (UNECE 2013c, ch. IV)

The model contains three levels describing the statistical process. When moving from the top level towards the lower levels, accuracy of the generic model improves and details are presented more exactly. Three levels of description and detail-refinement are

useful when integrating statistical data with metadata standards, harmonizing data processing systems, and assessing and improving process quality. (UNECE 2013c, p.3-4)

It is possible to identify several overarching processes that are applicable throughout the GSBPM. Mostly, these overarching processes relate to the management of specific issue such as quality, metadata, process data and knowledge. The quality and metadata management is specifically highlighted in this process model, with green color, for their importance in statistics production. (UNECE 2013c, p.5)

3.4.2 Specification of the GSBPM

Documentation

Documentation on the GSBPM-model describe specifically the overall process, each phase step-by-step and give examples for the common tasks that are typical for a phase. User may alter the order of the steps and remove unnecessary steps when describing own statistics process flow. UNECE documentation (2013b, p.24) encourage users to implement suggested practices and proceedings in a systematic way according to predetermined timetable.

The documentation also provide framework for the quality and metadata management reinforcing importance of these throughout the generic process model. Examples of activities that need to be done in the quality management are given although detailed instructions are lacking for organizations may use differing quality framework. (UNECE 2013b, p.10-26)

Information management

A challenge in the metadata management is to capture metadata immediately after it is created, to record the captured metadata and to pass it forward in the process. Part of this captured metadata associates to process while the other part to statistical data. Therefore, the GSIM model, that is a supplementary information model, was especially designed for the statistical data metadata management. (UNECE 2013b, p.25)

The aim of the quality management in the GSBPM is to control quality of statistical products and process. To render the quality management, new quality indicators for the

GSBPM model were introduced in 2015. New indicators were developed in a collaboration group having representatives from Canada, Italy, Turkey and Eurostat. (UNECE 2015, slide 2-3).

Quality indicators

The new collection of indicators contains totally 168 indicator that are linked to the process phases and sub-processes of the GSBPM. New quality indicators are defined for all process phases. Indicators are grouped according to the National Quality Assurance Frameworks, NQAF [9] that is a template used as a general structure in national quality frameworks.

At this first stage of indicator development work, the selection of indicators is limited and no exact formula is given for calculating indicators. So, these indicators need to be seen only as a reference model that will be improved in time. (UNECE 2015, slide 4-5).

In the new collection of indicators, quality is determined by broad sense. Some indicators measure the use of statistical methods and success of data collection, but also additional indicators are designed for the IT-monitoring and for general management. Additional indicators relate to the use of human resources, legal constraints, use of IT, standardization, documentation, test, archiving, delay in data transmission, collection costs and overall budget (UNECE 2015, slides 9-62).

Even though this information is valuable, these indicators does not belong to the context of this thesis. Most suitable indicators in this context are listed in the table 4.

[9] Template for National Quality Assurance Framework is developed to the request by the UN's Statistical Comission in 2010. The generic national quality assurance framework template was approved in 2012 and countries are encourage to use it. http://unstats.un.org/unsd/dnss/QualityNQAF/nqaf.aspx

Table 4. The list of new indicators developed in UNECE co-operation group

PHASE	QUALITY DIMENSION	INDICATOR
Design	Managing respondent burden	Percentage of questions used to collect information which will not be published (and motivation).
		Indirect evaluation of response burden: number of questions of the questionnaire
		Trend in respondent burden with respect to the previous iteration
	Methodological soundness	Extent to which the survey population matches the target population
		Timeliness of the frame: When was the frame last updated?
		Impact of coverage errors: Assess the likely impact of coverage error on key estimates.
		Key indicators for sample design (e.g. estimated size, expected/planned sampling errors for key variables, domains, costs)
	Soundness of implementation	When have the methodologies for subsequent phases (e.g. coding, E&I, data integration, estimatio) last been assessed?
Collect	Accuracy and realibility	The rate of over-coverage: The proportion of units accessible via the frame that do not belong to the target population (are out-of-scope).
		Rate of missing or suspicious stratification and classification variables;
		The sampling error
		Domain response rates; ;Unit nonresponse rate; item nonresponse rate; proxy rate
Process	Accuracy and realibility	Percentage of errors comes from identification and transformation of population, units or data items.
	Methodological soundness	Compliance rate of classifications of input data to the pre-determined standard international classification and national versions of international classification scheme
	Accuracy and realibility	Rate of actual errors: Identification of incorrect data (actual errors) in the processing stage - Missing, invalid or inconsistent entries or that point out data records that are actually in error.
		Imputation rate
		An indicator of an edit's effectiveness
		Edit failure rates. A sub-class of edits could be those designed to detect outlier observations.
		Rate of robustness of outliers for key variables. This indicator will measure the quality of outlier detection process
Analyse	Accuracy and realibility	Number of errors that were detected and had to be corrected

The list of indicators, in the table 4, is an extraction from the original 168 indicator list. The table shows that most of the indicators, in the sense of this thesis, relate to the dimensions “Accuracy and reliability” as well as “Methodological soundness”.

3.5 The GSIM and generic information models

The GSBPM model provides overall approach for describing statistics processes and information management, thus complementary framework, the GSIM, was designed to provide broader coverage of statistical metadata. As expressed in UNECE documentation (2013b, p.26) metadata

> should uniquely and formally define the content and links between objects and processes in the statistical information system.

In this chapter, the GSIM is presented similarly as the GSBPM previously. In the literature review, concern was put on the internationally set qualifications (see table 3. The qualifications for statistics compilation). These show that one of the most important qualifications are demand to provide information of procedures and statistical methods, to identify weaknesses in product quality and to regularly asses and validate source data, intermediate results and statistics outputs.

3.5.1 Overview

According to UNECE, the GSIM is the first internationally endorsed reference framework for statistical information (UNECE 2013d, p.3). It is a collection of standardized information objects, to be used in design and production of statistics (UNECE 2013a, ch. I). With the GSIM, statistician describes input and output-data related information fairly comprehensively (UNECE 2013b, ch. I).

Background of this development work is in the UNECEs' Strategic Vision that identified two major challenges: product challenge and process challenge. Conclusion to the met challenges, is to standardise characteristics that represent statistical information. (UNECE 2011b, p. 3, 7).

The figure 6 provides general view to the GSIM that is divided into two subsets: a conceptual model and implementation standards. The conceptual model covers identified information objects while implementation standards are used when introducing identified information objects into statistics production.

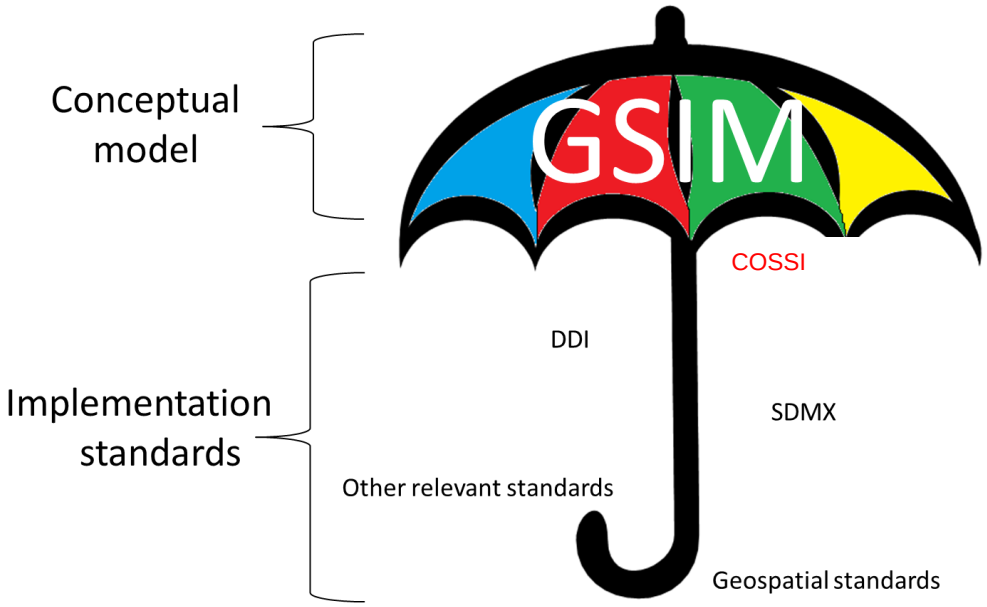


Figure 6. The GSIM conceptual model and its' implementation standards (UNECE, 2013e)

The model identifies around 110 information objects, such as data sets and variables of data, classifications, units, and population that define target group as well as parameters and rules that are used in data processing (UNECE 2013b, ch. Introduction; UNECE 2013d, p.4).

As stated in the GSIM description in the UNECE Statistics-wiki this information model

> does not provide any standard representation of its own, and is intended to be implemented using existing external standards and models, which support technical implementation.

UNECE (2013a, ch. Introduction) has named two information standard that are applicable for implementing objects into statistics production. These standards are the DDI Data Documentation Initiative[10], and the SDMX Statistical Data and Metadata Exchange[11].

[10] DDI is free international standard that is used to describe observation data in a lifecycle of this data. http://www.ddialliance.org/
[11] SDMX consist of technical standards including information model. This ISO standard is used to describe statistical data and its' metadata. https://sdmx.org/

Also the COSSI-model, designed in Statistics Finland, may be used as an implementation standard even though it is not originally presented in GSIM-conceptual model figure.

All these implementation standards meet the UNECE (2011b) requirements for an industry standard that are:

- To describe microdata—this typically address to data fields and observations
- To describe aggregated data – this address characteristics of population and recognising time series
- To be coherent by its' structure – all objects need to be modelled similarly
- Standard need to be human interpretable and machine actionable – this is often referred as "metadata driven business processes"

UNECE has also set sixteen core principles for the metadata management that belong to The Common Metadata Framework[12] that is part of the GSIM model. These principles are categorised in four groups: Metadata handling, Metadata authority, Relationship to Statistical Cycle and Users. These are overarching principles, so these need to be taken into account in designing and introducing a metadata system.
(UNECE 2013b, p. 26).

To summarise the Common Metadata Framework principles:

- Focus on overall statistical business process model and integrate metadata related work with process across an organisation
- Use active metadata that drive other processes and actions
- Reuse metadata where possible
- Minimise errors by entering once and updating in one place
- Capture metadata automatically (if possible) at their source
- Identify users and ensure that captured metadata creates value for users
- Ensure the availability and usability of metadata

(UNECE 2013b, p. 26)

[12] The Common Metadata Framework is developed through collective input on national and international organisations in 2004 to provide guidance for building internal metadata systems. http://www1.unece.org/stat/platform/display/metis/The+Common+Metadata+Framework

3.5.2 The SDMX Statistical Data and Metadata eXchange

The SDMX provides standard terminology for statistical data and metadata, content oriented guidelines for data transfer and standards for technology that may be applied in sub-processes. In 2009, this model was seen as a suitable method for data transmission between sub-processes inside a statistical institution, and for aggregated data transmission between two or more organizations (UNECE 2010, p.4, 7) so it was adapted to the GSIM.

The SDMX is widely supported by international organizations, such as the European Central Bank, Eurostat, OECD, IMF, the UN and the World Bank. These organizations sponsor development work thus this is one of the reasons why standards cover solemnly metadata describing aggregated data, a.k.a macrodata, in standard format (Praženka & Boško 2011, p. 2). Observation data, a.k.a microdata, is collected and manipulated internally in national statistical institutions while macrodata is provided for the use of national and international authorities.

3.5.3 The DDI Data Documentation Initiative

The DDI model is the result of international development work where the objective was to establish an international standard for describing survey data, such as registers, administrative data and questionnaire, and observational methods. (DDI Alliance 2016a). The DDI model is defined by the DDI Alliance group that consists of academic and research institutions.

The aim of the DDI is to offer a standardised approach to metadata for statisticians and researchers. As expressed by Arofan (2011) the DDI offers the standard information model that may be used to describe microdata and tabulated data, a.k.a aggregated data with descriptive metadata.

The DDI Alliance has two major development branches: the original DDI and the life-cycle based DDI. Both development branch may be used for describing statistical data yet having different capabilities. (Arofan 2011, p.2)

The DDI-model comprises of

- The lifecycle model, that is developed for documentation and data management purposes across entire lifecycle
- lifecycle containing technical specifications
- the original DDI, a.k.a the Codebook

(DDI Alliance 2016b)

The DDI3 Combined Life-Cycle model has certain similarities with the GSBPM even though development group did not include representatives from statistical institution. Similarities may be seen in the structure of models for both models contain common process phases. In the figure 7 is presented the process phases that belong to the Combined Life-Cycle model.

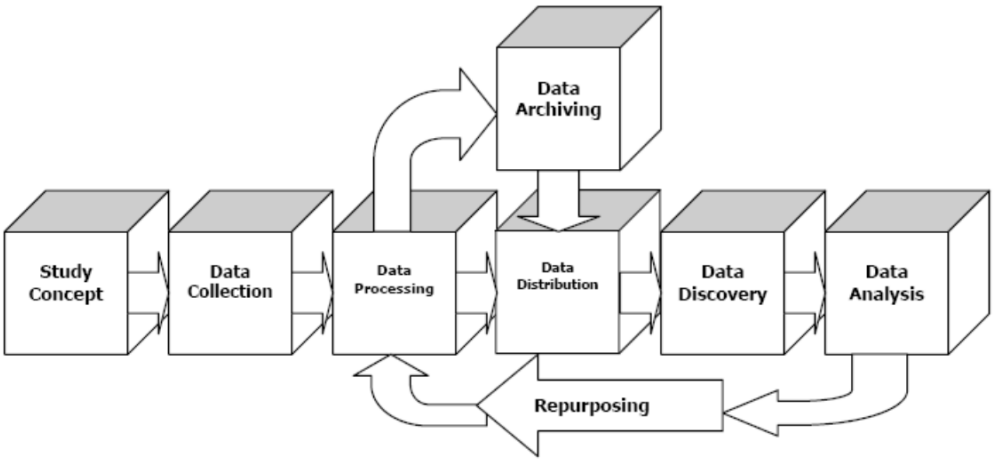


Figure 7. The DDI3 Combined Life-Cycle Model (Praženka, D. & Boško, P., 2011, p. 5)

As it may be noticed, the figure 7, overlapping phases with the GSBPM, are Data Collection, Processing, Analysis and Distribution. These similarities help adaption of the DDI standard in statistics production when moving towards metadata-driven survey design. The DDI model is based on the XML-standard that is machine-readable, so stored information may be used to drive processes and to support tasks in the life-cycle. (Arofan 2011, p.3). The DDI3 Combined Life-Cycle Model is particularly focused for describing information as it is created and utilised throughout the process model.

Beside the life-cycle model there has been growing interest in official statistics to the DDI-information model, ,namely the Codebook that may be linked with the SDMX-model and the GSBPM model (UNECE 2010, p. 4) to get full coverage to metadata in

process model. The Codebook may be suitable especially for documenting collected data (Arofan 2011, p.3)

3.5.4 The key areas in the DDI and the SDMX

The DDI Lifecycle model contains abilities that do not exist in the DDI Codebook. It offers standard way to describe survey instruments and repeated data collection cycles, enabling comparison across the cycles. It also recognises the re-use of metadata throughout process and offers multi-lingual standard. (Arofan 2011, p. 4).

According to the Praženka & Boško (2011) article both information models stand for similar artefact that are identifiable elements (ID-number), versionable elements, maintainable elements, notes may be attached to elements and xml technology. Both standards also describe data sets and their structures with common metadata components that are concepts, code lists, dimensions and attributes, measures, and the structure of aggregated data cubes (Arofan 2011, p.5)

These similarities render implementation of the models. Yet it need to be noted that there are differences that may hinder the adaption of the models.

3.5.5 Weaknesses in the DDI and the SDMX

The GSIM model is designed to be a framework so instructions for implementing it are very generic by nature.

Practical tests pointed out incompatibility of the SDMX for describing micro-data hence statistical institutions started to test combination of the DDI standard for micro-data and the SDMX standard for aggregated data. (UNECE 2010, p. 7).

The DDI Codebook was suggested by UNECE for describing statistical metadata. Some imperfections may be identified in it because it does not enable users to describe complex longitudinal or repeat-cross sectional surveys that have consecutive waves. It also supports only single language (Arofan 2011, p.2). This is not applicable standard in Finland for Statistics Finland’s statistics releases are bilingual.

Also Praženka & Boško (2011) identified low compliance of the models thus listing the major identified differences in the documentation. These are presented in the table 5.

Table 5. The major differences between the standardised models the SDMX and the DDI

SDMX-standard	**DDI-standard**
Used for macro-data – mainly in dissemination phase	Used for micro-data – in all GSBPM phase
handles better the large data matrixes	Handles better the observational data
Simpler structure than DDI – fewer compo-nents	Complex structure – large set of schemes
Includes process metadata	Includes both process metadata and archives metadata

These differences may hinder progress of the standard implementation. The DDI standard seem to have more benefits than the SDMX-standard but complex structure of the DDI model may retard its implementation in statistical organizations. Differences between the standards are also visualised in the figure 8 that is simple description of how the DDI and the SDMX standards may be used for describing data.

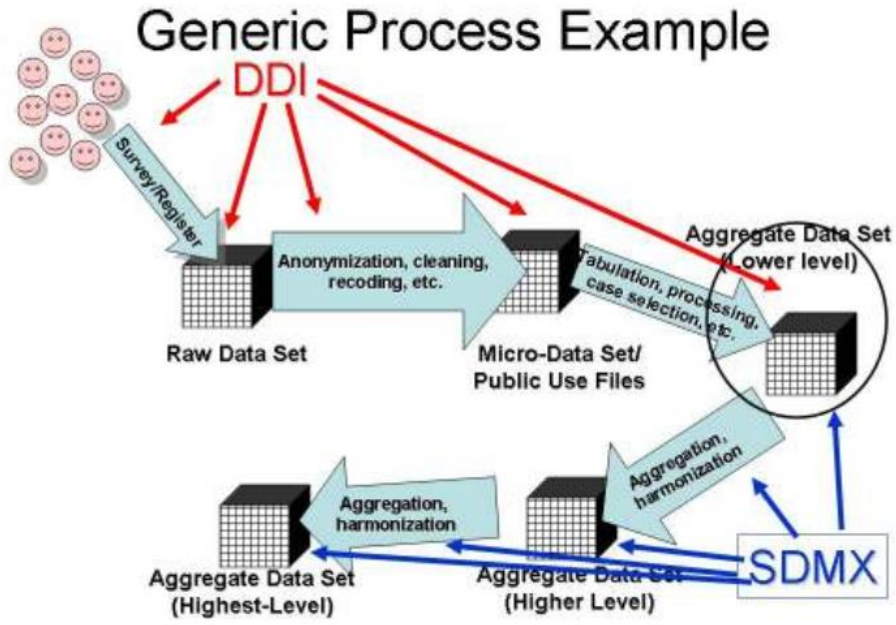


Figure 8. A generic example of the use of the DDI and the SDMX standards (Arofan 2011, p. 6)

As it may be discovered from the figure 9, above, the SDMX is mainly used after data is aggregated while the DDI is used in data collection and processing before aggregation.

3.5.6 The COSSI-model: Statistic Finland's information model for statistical data

The common metadata system in Statistic Finland is constructed based on the Common Structure of Statistical Information (COSSI)-model. It was designed in Statistics Finland by Heikki Rouhuvirta and Harri Lehtinen in 2007. The COSSI-model cover basic forms to organise statistical information and specifications for metadata that are required to describe it (Statistics Finland 2003). The model was implemented to the production of statistics after it was approved by director general. It is used as the common information model in Statistics Finland.

The COSSI-model contain metadata-elements that are used for describing the statistical data, such as concepts, variables and its' properties, links to classifications and content description, and for describing publication content and structure of statistical tables. Statistician stores information to the xml-database with an in-house developed application. Stored metadata is then utilised in the production and dissemination of statistics. The figure 9 provide a generic view to the COSSI-model components showing all planned metadata elements.

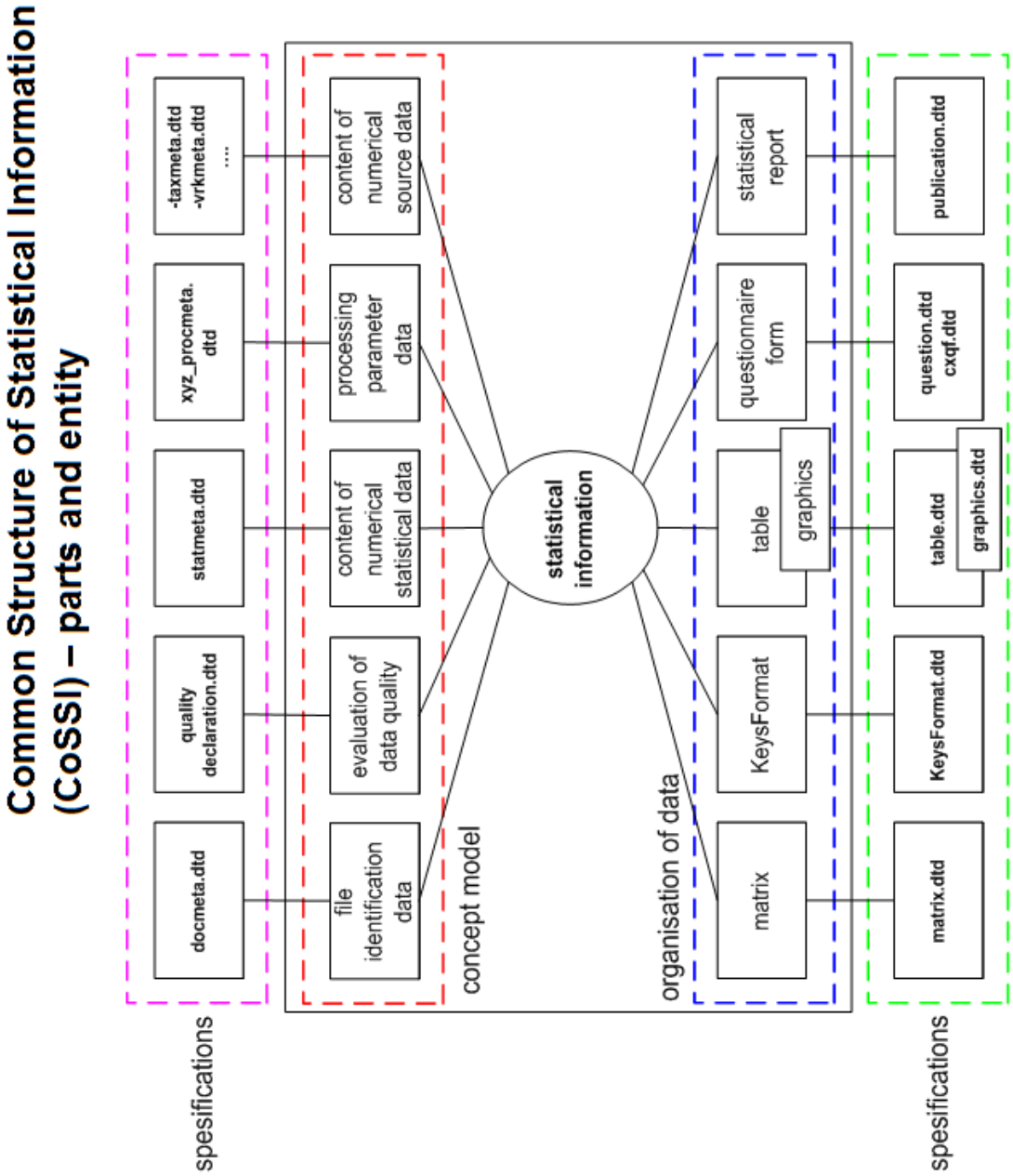


Figure 9. The Common Structure of Information Model designed in Statistics Finland (Rouhuvirta 2010)

The CoSSI model follow DTD-system that is based on modularity. This means that new modules may be inserted and unnecessary modules may be deleted from the COSSI-model. The COSSI-model is also based on the international standards such as CALS

Table Model, the XDF and the DublinCore[13]so it, as such, meets the international requirements. It also contains variable description in multilingual format so that this information may be utilised in statistics releases in Finland.

Content of the COSSI-information model has so far been enough for statistics production purposes. It enables developers to take new approach in the system design and programming without forcing developers to be dependent on certain IT-technology. Despite of this, it is important to expand the COSSI-model information content in order to meet the increased international requirements for more detailed information.

By contrast, the COSSI-model fulfils the UNECE requirements to be coherent by its structure and to be human interpretable and machine-readable. All of these requirements are met in the COSSI-model because it follows hierarchical xml-structure that is coherent and machine-readable. The COSSI-model uses the DTD-structure that is also easily human interpretable.

4 Current State Analysis

A strength and weakness (SW) -analysis was used to analyse current status of statistics production. The analysis started with an internal data collection, followed by the analysis of case studies and documents, and finally highlighting weaknesses and strengths in current practices.

The aim of internal data collection was to collect information on existing practices. This data collection was done with a web-questionnaire. Beside this, internal documents describing standardised actions and operations were investigated. As a result to this, a list of requirements for more detailed information was composed.

The composed list of requirements was then analysed in order to bring forward weaknesses and defects in current practices. In the final step in review, this list was compared with the framework qualification and the existing statistical information models and when necessary new requirements were inserted in the requirement list.

[13] More information about the Dublin Core and metadata namespaces, for describing information resources, are available in http://dublincore.org/specifications/

4.1 Overview of the Current State Analysis

Four data collection methods were used for collecting information for the current state analysis. Reason for using combination of methods was that it was important to get differing perspectives to current practices.

Firstly, we needed to understand current practices that are used to capture and to store information during statistics compilation, and to observe how well these practices acknowledge the requirements of the framework. It was also important to gather fresh ideas of how statistician utilise information in their statistics processes and what kind of lacks there is in the use of information.

Secondly, three case studies were investigated to observe practices that are already improved and standardised. Selected cases were the Administered Data Collection[14], the Generic Editing Model and the example of big data processing. All of these use elaborated information management methods.

Thirdly, Systematic Quality Audit reports were investigated and observations were compiled together for further analysis. Systematic Quality Auditing is a common method that is carried out annually at Statistics Finland. The aim is to audit specific statistics compilation process in order to outline 1) how statistics production is conducted, 2) what are practices in production and 3) what are weaknesses in production, introduced methods, resource-usage and IT-systems. (Statistics Finland 2016c).

Collected data was evaluated and a list of strengths and weaknesses was produced. Data collection for CSA is specifically presented in the chapter 4.2 Specifications of the data collection for the CSA. The CSA results are elaborately described in the chapter 4.3 Results of analysis. These findings are then combined with the findings from the literature review to form basis for an information model planning.

[14] The responsible unit, for processing The Administered data collection, is the Data Collection department in Statistics Finland. They use common practices to transfer administered data from supplier to target. Data transfer is managed with the uniform routing table. (Laurila, 2015, slides 13-14)

4.2 Specifications of the data collection for the CSA

This chapter specifies the data collection for the CSA and describes more closely how data was collected from several internal sources.

4.2.1 The web-questionnaire

The target population in this survey was 400 statistician and senior advisers working in three statistical departments and secretariat. The departments were Population and Social Statistics, Business Statistics, Economic and Environmental Statistics and Office of Director General.

Seventy four (74) employee name were randomly selected to the sample. An invitation for answering to the questionnaire was sent by email in spring 2016 with three weeks response time. One reminder message was sent in April to ensure the coverage of responses.

The opinion questionnaire was used to observe how well the GSBPM and metadata are known and how broadly these practices are used in statistics compilation. The questionnaire consisted of two background questions, ten option questions and five open questions. Only background questions were mandatory. The open questions were used for collecting examples of existing practices and examples of information that is expected for quality reports.

The questionnaire was constructed with the Digium Enterprise software that offers online responding. The questionnaire form is presented in the appendix 1.

4.2.2 Material for the case studies

Three case studies were selected to represent current, improved and standardised practices. These cases were the Administered Data Collection, the Generic Editing Model and the example of big data processing. All of these practices, use elaborated information management methods and are quite recently developed in Statistics Finland so they are not yet broadly implemented.

The three cases were selected because these were designed especially for statistics compilation. First case utilises the unified metadata system and the common practices to transfer and to process administrative data according to the pre-defined rules.

Second case shows the generic model of editing and imputing that was developed to systematically check and edit statistical data. This model provides comprehensive descriptions how to arrange process, what statistical methods to use and how to evaluate the results. It also contains accurately selected methods for collecting the quality information as an indicator during the editing process and for observing errors in data.

Third case provides an example of how to convert text-files into format that is more appropriate for calculation purposes. This case shows information that is needed for describing one actual process flow.

Case 1. The Administered Data Collection

Coincide with the web-questionnaire, all respondents were asked to send detail information about the use of metadata in their statistics compilation. Seven of all respondents sent additional information, of which the Administered Data Collection was selected for further investigation. This case offers applicable approach for importing various types of data to production.

The Administered Data Collection -documentation describes very thoroughly what kind of metadata need to be recorded before data transmission from supplier to Statistics Finland may be launched. The documentation covers also practices that are used for analysing imported data.

Case 2. The Generic Editing Model

The Generic Editing Model is an editing model designed and developed in Statistics Finland by Pauli Ollila and his project group. The Generic Editing Model was composed by putting together the best practices developed internationally and nationally.

The main sources for applicable practices and methods were the ESS[15], the EDIMBUS[16]- and the EUREDIT[17]-projects, Statistics Canada and Statistics Finland. In Statistic Finland also a web-questionnaire[18] (Statistics Finland/Ollila 2010a) was organised to investigate current practices that are used in data collection, editing and imputing.

The Generic Editing Model is designed to be used as a guideline for the steps and the actions that are needed for data editing and imputing. The model helps statistician to make correct decisions when processing so called raw [19]data, and making manual or automatic corrections to its' observations. The main phases in this model are 1) the analysis and the editing planning, 2) the editing process and 3) the evaluation phase. These phases, along their sub-steps, are shown in the figure 10.

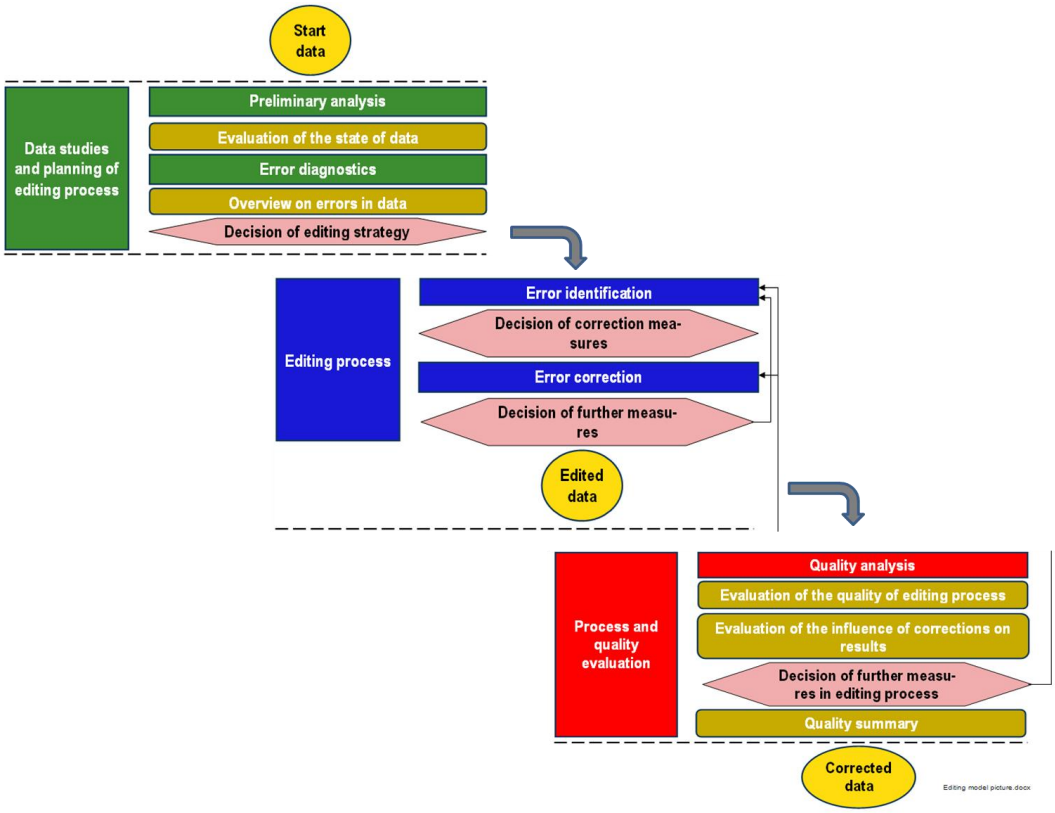


Figure 10. The main phases of the Generic Editing Model (Statistics Finland/Ollila 2012a)

[15] ESS, the European Statistical System
[16] EDIMBUS, the National Statistical Institutes of Italy (ISTAT), the Netherlands (CBS) and Switzerland (SFSO), where a Recommended Practices Manual (RPM) for Editing and Imputation in Cross-Sectional Business Surveys was developed
[17] EUREDIT, The Development and Evaluation of New Methods for Editing and Imputation –project coordinated by Office for National Statistics (ONS/Great Britain)
[18] Web-questionnaire contained 34 questions. Responses were received from 134 statistics so the response rate was 72 %. Questionnaire results were published in 2010 (Ollila, P., 2010b)
[19] A raw data –term is used to describe source data that is not yet manipulated in any way. The imported data is as it is in supplying system.

The process starts from raw data, in figure this is presented with the term “Start data” that is imported to the calculation system. The aim in the phase 1 is to study and to pre-analyze data in order to take an overview on typical errors in it. (Statistics Finland/Ollila 2012a, slide 2-3).

Data editing is done in the phase 2. This phase includes iterative actions that are used for error identification and correction. A basic rule is to follow the decisions made at the error identification phase. (Statistics Finland/Ollila 2012a, slide 4) The result of the phase 2 is so called edited data that has same observations as raw-data but does not include critical errors.

Finally, in the phase 3, the process and the quality are evaluated with the indicators that are calculated automatically during the process. Three types of indicators are automatically formed in order to summaries the data quality. These are 1) ”state of the data” indicators that are important estimates at population level and in relevant subgroups, 2) indicators revealing influence of the editing on results and 3) indicators in relation with previous results. (Statistics Finland/Ollila 2012a, slide 7). The result of the phase 3 is “Corrected data“, that contains same observations as the edited data.

According to the documentation, the measurement of data quality puts requirements for collected and stored metadata (Statistics Finland/Ollila 2012b, p. 3). Ollila (2012a) proposes that all editing model methods ought to be stored in a methodology bank that would include the existing concept library. In his proposal, structure of the methodology bank need to follow same grouping as in the editing model. These groups are presented in the table 6.

Table 6. Suggestion for the grouping of statistical methods

Measures describing the data	Refining the data	Search of value	Creating value
Realisation of unit view	Edit rules	Non-processed search of value	Non-processing creation of value
Realisation of listing view	Analytic processing	Defined search of value	Value with decision rule
Calculation of statistical measures	Macro level processing	Methodological search of value	Value with calculating statistics
Realisation of tabulation	Significance evaluation	Setting the value	Value with modelling
Realisation of analytical measures		Inputting value	Value with constraint application
Realisation of graphics		Setting values with written program lines	
		Values with predefined programs	

Ollila (2012a) amplifies that the methodology bank and the concept library ought to be easily available whenever statistician needs these for data processing, documentation and reporting.

Case 3. An example of the big data processing

The aim is to analyze, what kind of information is needed to describe an actual process and data.

At the moment there is a lot of fuss going around the term “big data”. From statistician point of view, interest is how to process and to analyze massive amount of data, millions of observations, according to statistical methods and pre-defined rules.

Statistical institutions use mainly two primary methods for collecting big data: 1) transfer data directly from supplier IT-system to receiver IT-system and 2) gather information from web-pages of an organization with a web-crawler[20] and an automated procedures.

In this example we take a look at the big-data processing that utilizes first data collection method. Data is automatically transferred from the external data warehouse on a monthly basis. Statisticians has developed program codes, with SAS[21]-software, that are used for processing the data. The process is demonstrated in the figure 11. It gives a simple view to the flow that contains

1. input files,
   in this example *.csv*-files,
2. one or several program codes, that contain processing rules
   here with names *collection* and *processing*,
3. results of a process in a form of a summary report or an output file,
   here output file:*sort3* and report: *SAS_report_laakeaineisto_tiedonkeruu*

[20] Web crawler is an internet bot that systematically browses the content of a web-page. (Wikipedia 2016)
[21] SAS software is software that allows statistician to access and manipulate data. http://www.sas.com/en_us/software/foundation.html

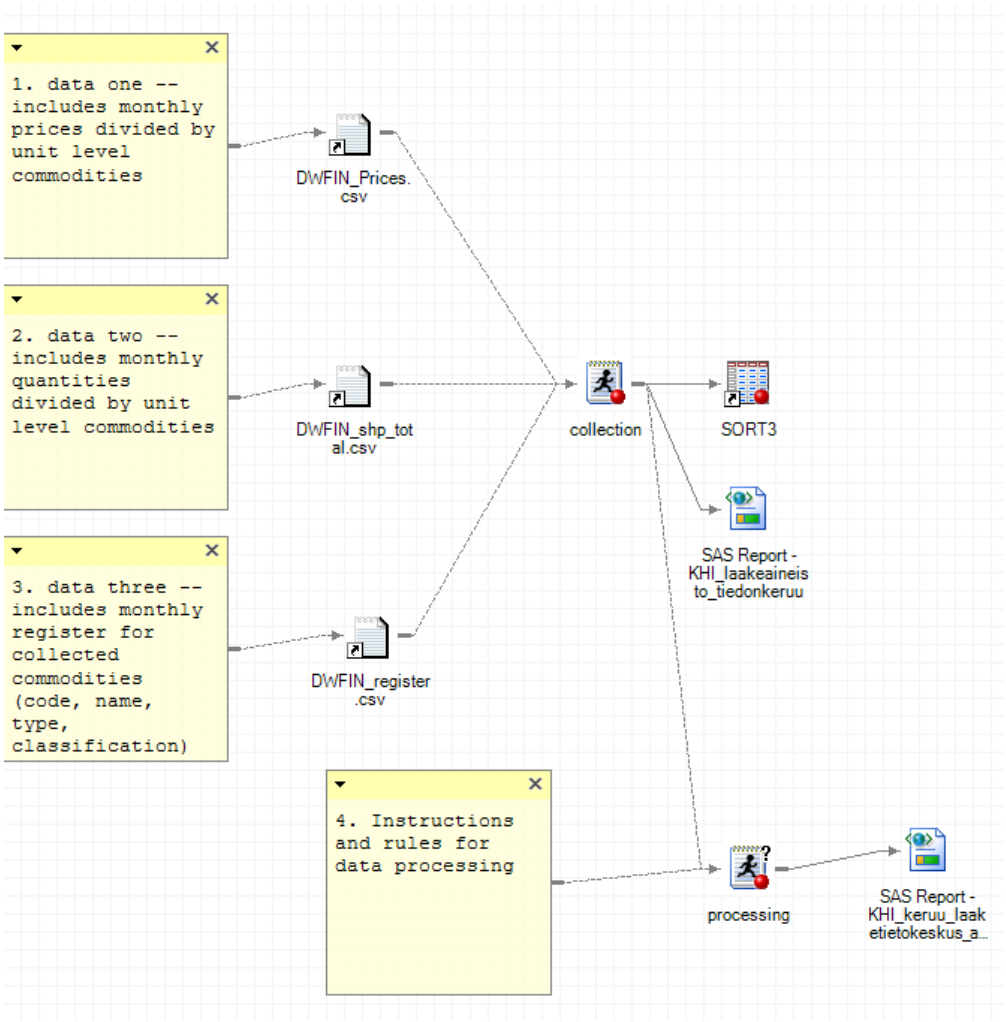


Figure 11. The demonstration of statistical data processing

The figure11 shows only first two steps that are used for importing and converting data into more adaptable format. Demonstrated process is very simple yet a lot of information is still needed to understand decisions, rules, formulas and data content.

In this example specific Data Descriptions [22]were retrieved from the common metadata system and combined with an input data in program called *Collection*. The data descriptions, in this case three separate data descriptions, were used for controlling the content of the source data, for defining the target table and for checking the content of an input file. Below, in the figure 12, is an example of data description showing portion of the stored information that is utilised in process above.

[22] The data description is a common method that is used to document content of statistical data. Documented data descriptions are store in the common metadata system in Statistics Finland and utilised especially in data dissemination.

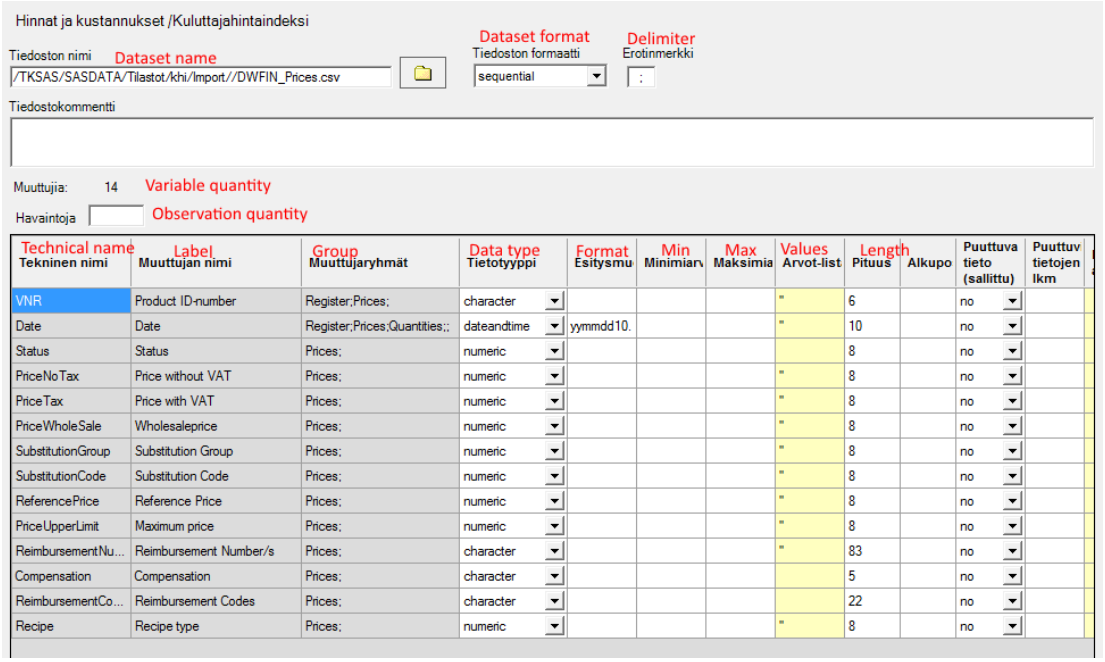


Figure 12. An example of data description that is used for the data transmission and data checking.

Similar kind of data descriptions are recorded in order to understand content of raw data. These data description follow principles and structure of the COSSI-model.

4.2.3 Material for analysis of Systematic Quality Audit reports

The aim of an auditing is to investigate one specific statistics compilation, to scan practices in use and to ensure compliance with international requirements and guidelines.

Random sampling was used to pick Systematic Quality Audit-reports from the collection of nearly 80 reports audited during the years 2012-2015, for this analysis. Selected reports were analysed to identify current practices and suggested development ideas for metadata management.

Following Systematic Quality Audit-reports were analysed; Consumer Price Indices, Prices of Dwellings in Housing Companies, Labour Force Survey, Indebtedness, Financial Statement Statistics on Credit Institutions.

4.3 The analysis results

According to the analysis results, several practices are in use to capture, to record and to store information but very little standardisation is node for capturing and recording the information. Only one standardised method, the Data Descriptions, is widely used for information management.

It was also noticed that the common metadata system seems to have enough information for accomplishing certain tasks but it does not provide information that meet the UNECE requirements, all of them. So, it is necessary to expand the content of the common metadata system with detailed process information, with the descriptions of implemented statistical methods and with the quality indicators.

Analysis reveal that first steps in information management and standardising practices was taken in 2005 when the common metadata system was introduced to the purposes of statistics dissemination. Complete introduction of the COSSI-model and the finalisation of the planned COSSI-metadata elements is still unaccomplished.

The CSA-results also show that especially process phase where data is manipulated lacks of standardisation. In 2012, was presented the model for standardised editing and imputing methods and for producing the key indicators for quality assessment. Only few statistics, from nearly 150 annually compiled statistics, has implemented these practices in to their production.

Next sub-chapters describe more specifically the analysis results and the conclusions.

4.3.1 Results of the web-questionnaire-analysis

This chapter presents results in more detail. All questions of the web questionnaire are covered one by one and the findings are concluded after each of them.

The response rate

Forty one (41) replies were received, so the final response rate was 55%. This response rate may be considered a good result. The replies were received quite evenly from different departments. Distribution of replies divided by department is shown in the table 7.

Table 7. The number and percentages of the responses divided by the department

	N	PctN
Population and Social Statistics	10	24.4%
Business Statistics	11	26.8%
Economic and Environmental Statistics	18	43.9%
Office of Director General	2	4.9%
TOTAL quantity of responses	**41**	**100,0%**

Second background question was about statistics production cycle the respondents are currently working with. It was allowed to select multiple options so the quantity of replies is higher than the quantity of received responses. The response rate is shown in the table 8 demonstrating that nearly 25 % of the respondents work with two or three production cycles.

Table 8. The number and percentage of the responses divided by statistics production cycle

	N	PctN
Monthly	20	37.7%
Quarterly	8	15.1%
Annual	17	32.1%
Other	8	15.1%
Quantity of given options	**53**	**100.0%**
TOTAL quantity of received responses	**41**	

The results of questions concerning the GSBPM

The questions four, five and six concerned the knowledge and the use of the GSBPM. The results show that awareness of the GSBPM is good as nearly 78 % knows the GSBPM. Yet, only 56 % has created process descriptions from one or several production phase based on this model. Share of the respondents, who has used the GSBPM as a guideline for describing process phases is presented in the table 9.

Table 9. The number and percentage of the respondents, who has used the GSBPM as guideline for describing process phases

	YES-answer		NO-answer		TOTAL
	N	**PctN**	**N**	**PctN**	**N**
Population and Social Statistics	7	17%	3	7%	10
Business Statistics	6	15%	5	12%	11
Economic and Environmental Statistics	9	22%	9	22%	18
Office of Director General	1	2%	1	2%	2
TOTAL quantity of received responses	23	**56%**	18	44%	**41**

***Conclusion 1**:* The table 9 prove that implementation of the GSBPM is underway in Statistics Finland although it is not yet comprehensively adapted. Only six respondents, approximately 15 % of all, has described all process phases according to the GSBPM.

After these basic questions respondents were led to the questions concerning the utilisation of metadata in statistics production. This issue was handled in questions seven and eight. Results show that only twelve of 39 respondents, 44% of all, has in some way described the process metadata.

The aim, in the question number eight was to reveal, if the respondents have at all described metadata, for example with excel, and what kind of information is described. This interests for researcher knows that there are statistical units, where own metadata practices are defined and in use. Twelve respondent gave positive answer whereof six offered more detailed explanation about the existing process metadata.

These replies were further analysed in order to understand and to get perspective to daily practices, and to get proposals for applicable metadata elements. Most of the transferrable ideas were received from the Administered data collection thus researcher decided to treat this as a case. The findings from Administered Data Collection practices-review are described in the chapter 4.3.2 Results of the case study-analysis

***Conclusion 2**:* There are differing ways to store metadata at statistical units. Some define metadata with the Ms Excel while others define important metadata in text documents or in programming code. Therefore uniform practices and structured information

model and -system is needed to enable statistician to record metadata, to observe it and to use stored metadata in their process.

The results of questions concerning data descriptions and classifications

The questions nine and ten covered the data descriptions and classifications. The objective was to observe how comprehensively respondents describe their statistical data, and utilise existing data descriptions and classifications at the moment. Based on the replies, nearly 70% of the respondents have used common methods to create and to store metadata. The results are shown in the table 10.

Table 10. The number and percentage of the respondents who use common methods for describing data and classifications

	YES-answer	**NO**	**TOTAL**	**Yes-answers of total**
	N	**N**	**N**	**PctN**
have created the data descriptions	28	13	41	68 %
have created the classifications	25	15	40	63 %
recognise process phases where these descriptions may be utilised	35	5	40	85 %

Conclusion 3*:* The figures in table 10 confirm assumption that data descriptions and classifications are well-known and these common practices are used at statistical departments.

The process phases, where data descriptions and classifications are currently used, were analysed from the replies given to the questions 11, 12 and 13. The aim was to identify those phases where the respondents are currently using data descriptions and classifications. The results reveal that there is variation when replies are divided by process phase. These results are represented in the table 11 and in the figure 13.

Table 11. The number and percentages of respondents who use data descriptions in production, divided by the process phase

	YES-answer	NO	Yes-answers of total
Process Phase	**N**	**N**	**PctN**
Collect	16	35	46 %
Process	15	35	43 %
Analyse	17	35	49 %
Publish	23	35	66 %
Other	5	35	14 %
Do not know	4	35	11 %

These results point out that around 46 % of the respondents use data descriptions in Collect-phase, nearly 43% use in Process-phase, while 49% use data descriptions in Analysing-phases. Best scores are received for Publish-phase where over 66% of the respondents use variable specific information that is stored as data descriptions to the common metadata system.

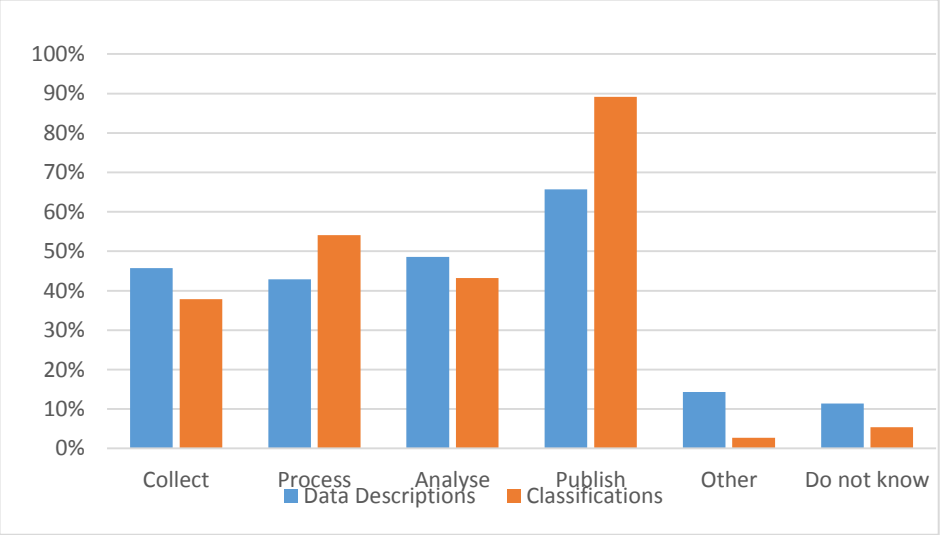


Figure 13. Process phases where the respondents utilise data descriptions and classifications by process phase

Similar analysis was carried out for the use of classifications. As we may notice from the figure 13 the use of classifications give similar results as the use of data. Best scores are received for Publish-phase, the process phase where unified practices were introduced in 2009. Seven respondents use data descriptions and classifications throughout the process.

***Conclusion 4**:* The findings, so far, may be summarised shortly. Awareness of the GSBPM, data descriptions and classifications is good. Yet common practices need to be developed and trained in order to implement this standardised approach throughout the statistical departments and units.

Finally, before the open questions, only one question (number 14) concerned structure of the COSSI –model. The aim in this question was to examine how many of the respondents already know the COSSI model. The assumption was that if you understand content of an information model, then you may understand possible ways to utilise this information in your process. Only seven out of 40 respondents (18%) answered.

***Conclusion 5**:* Awareness of the COSSI-model is moderate, so more training is needed in order to ensure that statistician clearly understand utilities of the COSSI model. A comprehensive understanding is needed especially if statistician works as a process developer or a senior adviser. More you understand content of common information models and standards, more easily you identify where this information may be captured and utilised.

The results of open questions

The questions 15 to 17, were open questions where the aim was to survey current practices in statistics production, to reveal demands for more detailed information and to show weaknesses in existing practices. These replies were thoroughly analysed since these provide most of new ideas for an initial model design. All received replies were listed to sum up given ideas and to calculate popularity of an individual idea.

The respondents were asked, in the question 15, to provide examples of developed practices for metadata management. These responses confirm previous results that are shown in the tables 10 and 11. Although good practices are available for statistician to utilise, only few department utilise these at daily work.

In the questions 16 and 17, the respondents were asked to share their views about quality and process information that ought to be collected and produced. Statistician work regularly with this challenge when they need to report results to Eurostat. They need to have good grip on issues regarding data and process quality. The respondents were asked to give two to five examples and to categorise their ideas by process phase.

Twenty seven (27) separate ideas were received for the question 16. Only one idea, “the effects of editing and imputing to the results”, was given several times. This is obvious result for quality of statistics depend on introduced editing and imputing methods.

Two ideas, “the history of made corrections by observation” and “the number of corrected observations” were given for both questions. This means that some see these as exiting practice while the others have not taken this into production.

Forty six (46) separate ides were provided to the question 17, where the respondents were asked to specify what kind of status and quality reports they would like to get from statistics production. Some ideas got more votes than the others. Next list presents the most popular ideas:

- need to know schedule of designed and executed process and its’ phases, need to get summary report from comparing two or more data,
- need to list automatically edited observations,
- need to list observations that were identified to be erroneous,
- need to identify primary variables in data,
- need to store run-time information of process such as performance and errors in the process execution
- need to flag manually and automatically edited observations,
- need to flag outlier observations,
- need to describe effects of editing and imputing to the results,
- need to understand data content such as distribution of variable,
- need to store what changes were done to data
- need to describe the success of data collection with indicators such as response rate, quantity of the observations in the data

***Conclusions 6**:* The results show that statistician are willing to share their current practices (question 16) and they know very well their needs for more detailed information (question 17). This confirm assumption that more information is needed for describing data processing, data, introduced statistical methods, rules, boundary values and most of all about quality of results. The total list of replies is comprehensive offering many ideas for building an initial proposal.

4.3.2 Results of the case study-analysis

This chapter presents the case study –analysis results in more detail. These three cases were selected because they are important examples as renewal of statistics compilation. The first case utilises the unified metadata system and the common practices for transferring administered data while the second one handles the generic editing and imputing methods that were developed for statistical data processing. Third one gives the example of how text-files are converted into more suitable format.

Case 1. The results of the Administered Data Collection – analysis

The Administered Data Collection documentation (Statistics Finland/Data Collection, 2016) highlights complementary metadata –elements that are Order, Supplier and Receiver. The aim is to use these elements beside the common data descriptions.

However, this complementary metadata content is very simple and minimal, it covers all information that is needed for data routing from supplier to receiver. Examples of the additional metadata elements are listed below

- filename and location
- target statistics name (=owner)
- information that is used for routing data from source to target
- information describing receiver such as server, transmission protocol, target folder name, receivers' email-address
- information describing supplier such as supplier ID, server, import protocol, source folder

These metadata elements have already been defined and stored to the common metadata system in Statistics Finland. This means that content of the common metadata system is expanded with this supplementary information.

Conclusion 7*:* The described practices in this working method are useful, yet metadata enlargement is quite limited. This does not meet statisticians' requirements that are details describing order, agreement and information about supplier.

Conclusion 8*:* Practices that are used for data analysis are useful, yet slight improvements ought to be done to achieve proper functionality. At the moment pre-analysis is

carried out for all variables in data. This approach consumes resources too much. So, primary variables need to be defined to data descriptions in order to use this information for limiting calculations.

Case 2. The Generic Editing Model results

All material of the general editing model is very well written so it was easy to identify its' strength, the indicators, that need to be noticed in an initial model planning. The documentation (Statistics Finland/Ollila 2012b) provide the lists of indicators for

- quality reports (table 22),
- analysis of raw data (table 23),
- controlling of editing process (table 24),
- corrected data quality analysis (table 25).

These lists were analysed and the final list, that contain sixty-seven (67) individual indicators, was composed. The documentation recommends to produce thirty five (35) of these indicators, while 32 indicators are discretionary and are advised to produce when necessary. Examples of indicators divided by topic are presented in the table 12.

Table 12. Examples of proposed indicators by topic

TOPIC	**example of preferred indicator**
Indicators that describe raw data	• weighted variable response rate, • variable response, • weighted variable response rate in proportion to auxiliary variable, • proportion of complete responses
Indicators that are used for an error identification	• detection rate of error identification, • overall rate of error identification, • proportion of detection rate by variable
Indicators that are used for measuring the editing and imputing actions	• rate of editing, • net rate of error corrections, • rate of deleted values, • weighted proportion of corrections

Indicators that are used for evaluation the quality of results	• variable response rate after error correction, • weighted variable response rate after error correction, • weighted variable response rate after error correction in proportion to variable values, • share of observations having item non-responses after imputations

Conclusion 9*:* All these indicators need to be taken into account in an initial proposal design. Measures describing and refining data are essential for quality reports. The web-questionnaire, that was organised in 2010 (Statistics Finland/Ollila 2010a), provided a view to information that methodologists value most.

Conclusion 10*:* A methodology bank that contains information about statistical methods and concepts need to be designed and implemented in Statistics Finland. The methodology bank, if founded and updated by official statistical office, is useful for internal users but also for external users, like analysts, students and a teachers.

Case 3. The big data processing –Evaluation made based on own experiences

Statistics are increasingly improving collection methods to acquire comprehensive data for statistics compilation. What interests most is how collected data is processed and what kind of information is needed to automate data manipulation process and to describe data and quality of results.

Information that is used for data processing is partly IT-related and partly content related metadata. Following information is needed to run process and to understand data and its' quality:

- *IT-related:*
  - name and location of source data files,
  - process container [23] and program-codes that are used in process
  - schedule of process step execution
  - monitoring information about process progress

[23] SAS Enterprise Guide is used in Statistics Finland for the process flows that statistician and methodologists develop. Developed processes are stored in a form of project that contain programs, results, output lists and other components needed to run the process (High & Miller 2007).

- *content related*:
  - detailed information about input data origin such as data collection method, presented questions, population, sample size
  - content of input data such as list of variables, information about variable type, format and classifications)
- *quality related*:
  - key figures such as number of missing values, number of non-complete observations, distribution of observations

***Conclusion 11**:* At the moment IT-related and content related information is not yet stored in unified and structured form, instead statisticians store this kind of information with a varying methods. Therefore unified method, to store and to utilise process and data related information, is needed. The common metadata system is too limited, so it need to be enhanced in order to let statistician record specific information systematically.

4.3.3 Results of Systematic Quality Audit-report analysis

All analysed reports were quite general especially when observed the provided development ideas in the context of the GSBPM and metadata. Although, few new requirements may be deduced from the development ideas:

1. need to describe more metadata from process and its' phases
2. need to describe detailed metadata of processed data
3. need to implement the GSBPM and metadata capture to process
4. need to define indicators that describe results and process quality
5. need to implement these indicators to statistics production
6. need to describe, more closely, data content
7. need to describe detailed information about data validation, compilation methods and confidence interval

***Conclusion 12**:* These results again confirm the need for more detailed information. Same perspectives, as was seen in the questionnaire analysis are presented here.

All conclusions are summarised in the appendix 2.

5 The design and assessment of an initial proposal

The CSA was designed and carried out to examine existing practices in information management and to gather requirements for more detailed information. When these results are combined with the findings of literature review, the basis for an initial proposal design is initialized. Next we take a recap on the findings so far and describe the design process more closely.

5.1 The summary of current state analysis and literature review -results

The aim was to design an initial model that takes account information elements that were pointed out in the literature or received in the CSA.

Here is a short summary to the current state analysis- results:

- need to identify and describe statistical data more thoroughly
- need to describe process flow and give detailed descriptions for process steps and actions
- need to describe and produce quality indicators
- need to produce reports describing data content, success of process and quality of data
- need to describe methods that were used for data processing

The model design was done by paying attention to two perspectives: data and process. Data related metadata was divided further into two groups, microlevel and macrolevel. This was seen to be important because requirements for metadata may differ slightly depending on data content. Also Open Data Foundation/Arofan has taken this approach in their example for generic process (figure 9).

To clarify differences between microlevel and macrolevel data, the microlevel data is an observation matrix where single survey unit is represented in a row while measured variables are shown in columns. Whereas macrolevel data shows results in aggregated format where observations are grouped together representing summarized information.

When observed process flow, the order of the process steps is important. Example of the GSBPM structure is presented in the figure 5. The GSBPM-figure shows that statistical process flow has direction from start to end. It also need to be noted that it may have simultaneously parallel or iterative sub-flows. Hence, process structure need to be seen two-dimensional having vertical and horizontal direction.

The generic business process model is divided into several levels that may be divided into phases and steps. The highest level 0 (zero) ought to offer generic description of process while lower level descriptions specify and provide detailed information about objective of step, implemented methods, input and output data and processing parameters. Deeper we move vertically in process flow structure, more detailed descriptions of specific step, or action, and data is needed. Lower level process may receive input information from higher level process and vice versa.

5.2 Design of new metadata elements

At first, all CSA- and literature review findings were put together to create a comprehensive requirement list that is used as a starting point in designing an initial proposal.

The items in the requirement list were then converted into information fragments that belong to specific metadata element. Idea was to link information fragments with metadata elements so that concepts describing the same phenomenon go to same group. Lets' look with an example, in the table 13 how proposals were transformed into an information and linked to metadata elements.

Table 13. An example of how to transform received proposals into information, and grouping these in selected metadata groups

Received proposal	**Identified requirement**	**Suggested information fragment**	**Metadata element-group**
It is important to identify primary variable	Primary variables need to be recorded in variable specific descriptions	Primary_variable	DataDescription
We need to know what was the data collection method	Data collection-method need to be recorded to the data description	Collection_method	DataContent
What is the program code that is executed in certain step	Name of the program code and location need to be recorded in process description	Program_name Program_location	ProcessingMeta

The table 13, above gives three examples of the identification process. The first proposal in the table is “the need to identify primary variables”. Primary variable –information is needed to recognize most important variables of data that may contain dozens or even more variables. Requirement is clear. Suggested information fragment here is *primary_variable.* Suggestion does not pay attention to the content, nor format, of this new information fragment, but only shows the need for this. Then this suggestion, *primary_variable*¸ was considered closely in order to be able to link it with suitable metadata element-group; data or process. Bond with data is tighter so this new information was linked to metadata element “Data Description”.

Similarly all other proposals and requirements were examined. Finally, a list of new metadata elements and their information fragments was composed. Next we take a look at this list and what is structure of the initial proposal.

5.3 Introduction to the initial proposal

Primary objective, in the design, was to design a plan of an information and metadata elements, existing or new, that are needed for describing statistical data and process. Second objective was to complement this plan with monitoring information that indicates success of the process and the quality of results.

The designed plan, that contains new and existing metadata elements, is summarized in following list. All elements have a short specification and suggestions of information fragments that belong to the group. These metadata groups are also presented in the figure 14 below. The metadata element name is shown in the list with italic-font[24] and the suggested information fragments are listed below each element, while brackets are used to show whether metadata element is new or existing one.

Metadata describing the data

- *DataCollection (new metadata element)*
  Describes data collection method and measures success of data collection
    - Data collection method (e.g. web-questionnaire, interview, register).
    - Description of target population, sample and questions
    - Key indicators: Response and non-response rate, quantity of responses and reminded respondents, accumulation of responses by primary variable
- *DataDescriptions (existing metadata element)*
  Describes statistical data; microlevel data or macrolevel data
    - Description of detailed variable information such as variable name, presentation format, measure, composition rule, concept and classification
    - Describes general information about the data such as owner of the data, content and location
- *DataContent (existing metadata element)*
  Describes the data content in general
    - dataset name, unique identifier, source of data, user of the data)
- *FileCatalogue (new metadata element)*
  Lists all datasets in Statistics Finland including raw data, edited data, final data, permanently stored macrolevel data and statistical tables
    - file name, location and format, delimiter of the variables
- *DataQualityMeta (new metadata element)*
  Describes quality of data and quality analysis methods
    - Description of life cycle impact-analysis and its' results, how data editing and imputing affects the quality of data, impact of coverage error

[24] From now on in this thesis designed metadata-elements are presented with italic font.

    - Description of comparison methods and their results, and checking methods
    - Key indicators for quality analysis: aggregated results, distribution of variable by classifications, timeliness of frame, imputation and editing rate
- *VariableQualityMeta (new metadata element)*
  Describes quality of a variable
    - key figures by variable: monthly and annual change, quantity, accumulative quantity, standard error, share of imputed values, variation, standard error, confidence interval, response rate, correction rate, item non-response rate
    - description of methods that were used to identify erroneous values
- *ObservationMeta (new metadata element)*
  Describes quality of an observation
    - Key figures describing the quality of an observation: net imputation rate, correction rate, response rate, exclusion rate
    - Description of disclosure control and quality assurance methods used for an observation
- *Classification (existing metadata element)*
  Describes classification content, lists accepted codes with verbal descriptions, provides additional information by code-value. Classifications are linked with the variables defined in the data description

Metadata describing process

- *ProcessingMeta (partly new metadata element)*
  Describes metadata that relates to process phase and process step by production cycle
    - Verbal and graphical description of the process flow
    - Name and location of the program code with the link to the process step
    - Planned production schedule and realized execution date and turnaround time
    - Description of checking methods that are used during and after the execution of specific step

- *MonitoringMeta (new metadata element)*
  Describes metadata relating to the success of the process
  - Description of the observed challenges by production cycle
  - Key figures: observation count by process step, conclusiveness of the schedule, check of the input data, rate of actual errors (missing, invalid or inconsistent recordings)
- *StatMethodMeta (new metadata element)*
  Describes statistical methods that were used for data processing, and validity and correctness of the sample and data collection
  - Key indicators: sample design, coverage, missing and suspicious variables, sampling error
  - Verbal descriptions of processing rules, and data editing and imputation

Complementary information that may be needed in process or when developing the process

- *Operational Guidance and Planning System STOJ (existing information)* STOJ is used to store information about the timing of data collection and publishing. This information may be used as a trigger in process automation
- *Working Instructions (existing information)*
  These are composed for each statistical process. Instructions contain verbal and visual presentation of statistical process by phase offering detailed guidelines for statistician who carries out daily process step by step.
- *IT-architecture (new metadata element)*
  Describes infrastructure of whole IT-system. This information may be utilised in data processing
  - Servers, databases and folder structure used for storing the data
  - Tools and software used in data processing
- *Methodological library (new metadata element)*
  Contains concept library, describes statistical methods and instructs how to use those methods, describes how to measure impact of data manipulation
- *Background information (new metadata element)*
  Describing the statistical specific information
  - Legislation, instructions, manuals
  - Recommendations

The figure 14, provides graphical presentation of the designed metadata elements and complementary information. The connection links (arrows) show logical relationship between the metadata elements. Starting point here is the metadata element called *DataCollection*, from where the first arrow is drawn to the metadata element *Classification* that provides classification related information that may be used in data collection.

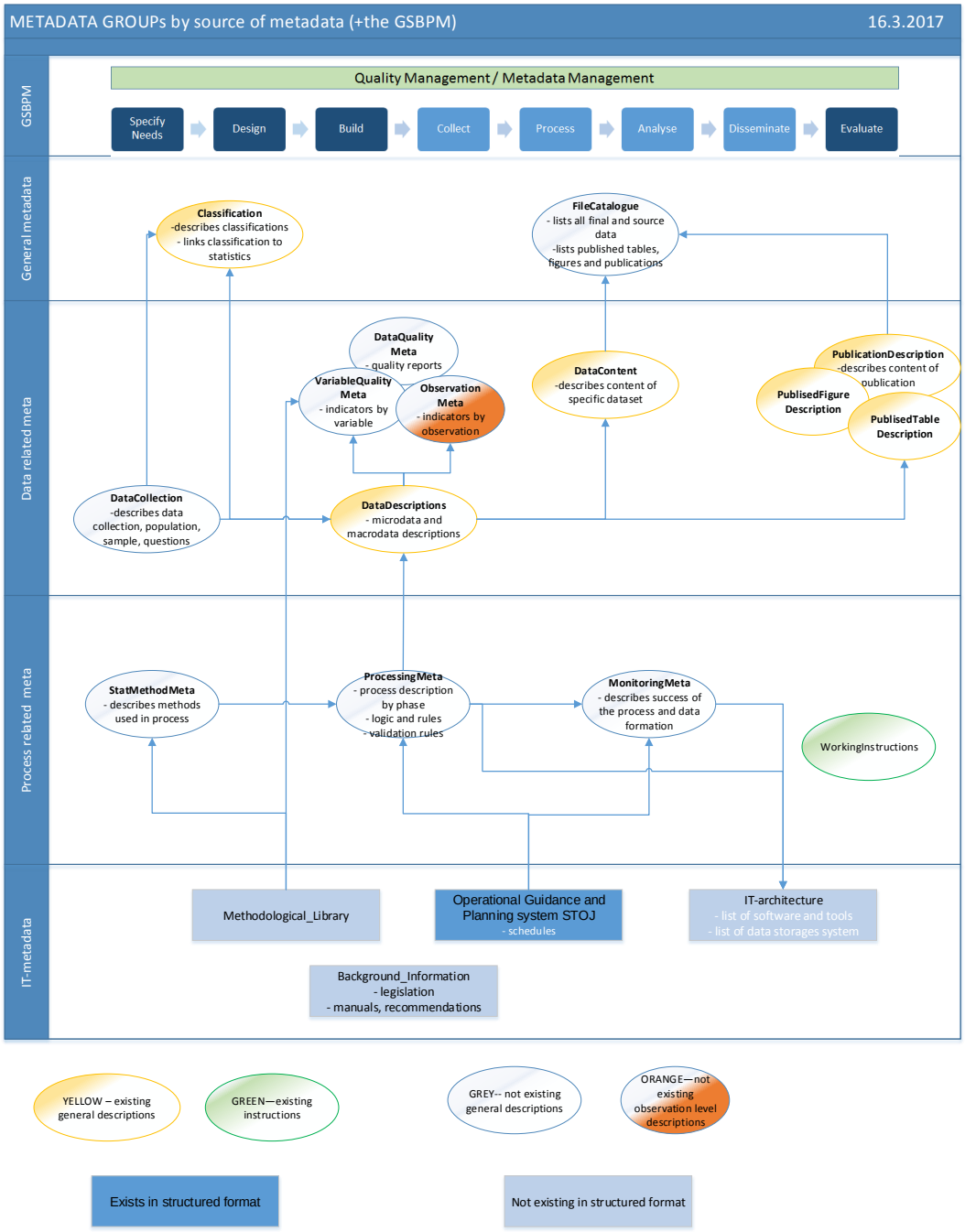


Figure 14. The graphical summary of the initial proposal

It may be easier to understand links between the metadata elements with an example. In the figure 14 *DataCollection* is in relation with *Classification* and *DataDescriptions*. Classifications are needed in data collection to describe actual values of classification variable. For example, Individual Consumption- classification and its' descriptions are needed to identify and distinguish the products and services in the data collection. Same way, the data descriptions are needed in the data collection to identify the variables of collected data. For example, in Consumer Price Index-data collection, it is important to separate a unit-price from a package price. This is not possible if only data values are observed, instead also specific data description that describe the variables more specifically are needed.

The idea, in the initial proposal, was also to orientate metadata elements with the GSBPM-model and its' process phases. It was not possible to illustrate this in the figure 14 because one metadata element may be joined with one or more process phase. Hence, the figure 15 was drawn to illustrate how the metadata elements are taken as an input information to the statistics production, in the process phase Collection.

The figure starts from the box *FileCatalogue*, on the left that contains information about the statistical files in Statistics Finland. The files are linked to the *DataContent* descriptions that more elaborately describe content of one specific data. Then arrows point to the *DataDescriptions*, *Classification* and *DataCollection*. All these metadata elements provide detailed information about the dataset that contain the variables and the observations in the structured format.

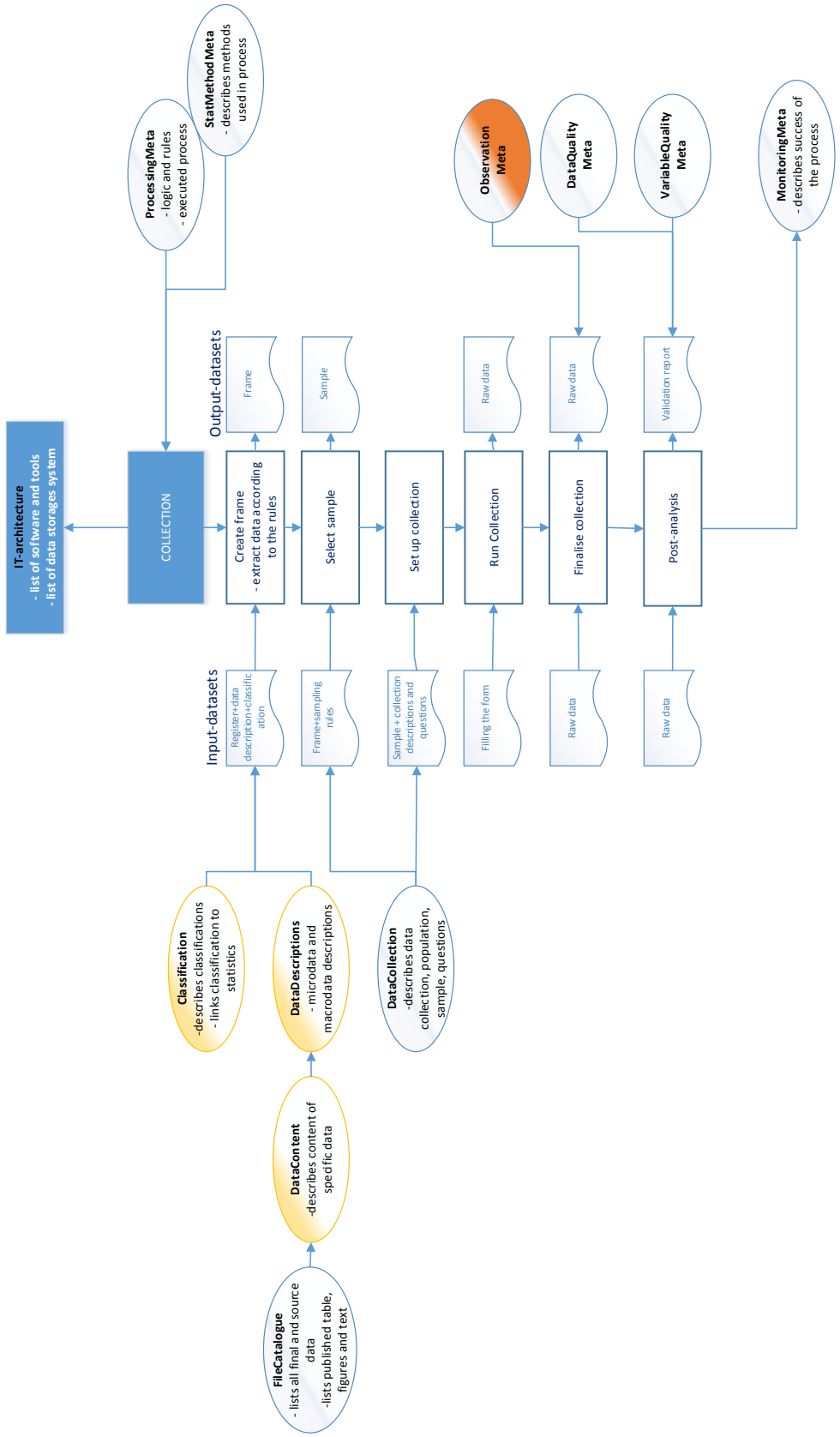


Figure 15. An example of the metadata elements used as an input or output information in the phase Data Collection in GSBPM

The input datasets are taken into the statistical process together with the metadata elements describing process: processing rules (*ProcessingMeta*) and methodological descriptions (*StatMethodMeta*). The figure illustrates metadata elements that relate with the outputs of process, such as *DataQualityMeta* and *VariableQualityMeta*-groups. This kind of information is needed in order to describe quality of the results as well as success of process (*MonitoringMeta*).

This initial model was presented and assessed in the focus group meeting in 16th March 2017. Next chapter “5.4. Assessment of the initial proposal explains how the assessment was done and summarises the received feedback.

5.4 Assessment of the initial proposal

5.4.1 The composition of focus groups and working methods of the assessment

The initial proposal was presented to the focus group in March 2017. The aim of the meeting was to assess the proposal and get feedback for finalisation of the information model. Successful assessment of the proposal is dependent on the competence of evaluators, so most of the participants were invited to the focus group. The focus group was composed of participants from the statistics producing departments, the metadata department and the IT-department.

The evaluation meeting started with a presentation that outlined the aim of this work and presented the initial proposal. Printed copies of the initial proposal, figures 14 and 15 with the list of the metadata elements (p. 58-60), were handed to the participants before the presentation.

After this, participants discussed shortly of the drafted metadata elements and especially their information fragments. At the end of the meeting participants filled the ten question-questionnaire (see appendix 3).

Finally thesis mentor, Heikki Rouhuvirta, gave his opinion to the initial proposal, outlining its benefits and shortcomings.

5.4.2 The assessment results

Totally nine responses were received to the assessment questionnaire. All responses were recorded and analysed further. Following table 14 lists identified strengths and weaknesses of the initial proposal, and provides suggestions for improving the model further.

Table 14. The remarks and suggestions for improvement of the information model

FEATURE	**Received comments**
Strengths	Proposal includes exhaustive collection of metadata elements
	Great and important work that is a good starting point for the SF's own information model-design
	Comprehensively encompasses different process phases of the GSBPM
	Model is process- and statistics oriented
Weakness	Proposal need to be clarified
	Concepts and vocabulary are missing
	Includes lot of new requirements for capturing the information
	Point of view is separate dataset, not the data warehouse that is current trend in SF
	The links between metadata elements is not clear
	International GSIM/ESQR/LIM-standards are not thoroughly taken into account in the proposal
	Arrows (links) are pointing here and there without clear meaning of their direction
Suggestion	Most of the metadata should be automatically produced

Beside these results, Heikki Rouhuvirta passed his view to the initial proposal. He suggested to reconsider the information model and take slightly different perspective in finalisation of the model. Reasons for this suggestion was that the GSBPM process did not link sufficiently to the planned metadata elements. All planned metadata elements may

be split into smaller information fragments, but link with the GSBPM was unaccomplished. Hence, further consideration was needed in order to identify essential information.

The final step, in the model planning was to re-design the information model based on the acquired experiences. Most important finding was that the link between the GSBPM-phases and the metadata elements was non-existent. The main focus in finalisation of the information model was put on establishing a link between process and the designed metadata elements. Also the content of metadata elements needed thorough walkthrough and improvement.

Summary of the final solution is demonstrated in next chapter 6. Final solution

6 Final solution

Focus in the re-design, was set to process metadata that is captured and used in the GSBPM process, and content of metadata elements. So the main questions for the model finalization were:

- What process information is important and need to be stored in metadata system?
- How to link, process metadata with existing statistical data and data descriptions?
- Is there need to expand existing data descriptions with new metadata?
- What is structure of final information model?

In order to see the essence of these questions, we needed to once more look back in the thesis process and draw conclusions from findings and suggestions.

6.1 Conclusions so far

So far we have observed current practices in Statistics Finland, done a literature review, drafted the initial proposal and also assessed it in focus group meeting. Here are the conclusions deduced from the preceding work:

1. statistician need detailed information of statistical data and process where this data is manipulated
    a. it seems that existing metadata system in SF is not sufficient to fill this demand, so it need to be expanded. The proposed information model for statistical data need slight amendments while information model for process metadata need to be created from scratch.
    b. it seems that international information models and instructions, like the GSIM, the DDI, the SDMX, handle metadata that concern statistical data and data delivery but not process. So it seems most reasonable to construct own information model and link it with the existing international standards.
2. metadata administrators and IT-architects need an information model that is understandable and may be introduced in to production
    a. the GSIM offers an information model that follows object –model, whereas hierarchical structure is used in the common metadata system at SF. It seems that most sound solution is to design an information model that follows hierarchical structure
3. statistician need centralised library for statistical methods in order to understand methods more precisely and to take these rules and formulas into process
    a. the editing model-project group suggested this at the beginning of this decade but still its implementation is unfinished. So it need to be created

6.2 Presentation of the final information model

6.2.1 Overview to the final information model structure

Next, we will take a look at the final information model structure. As noted in the previous chapter, most sound solution is to design an information model that follows the hierarchical structure because the common metadata system is based on it.

The hierarchical structure means that there may be, and most commonly is, multiple levels of information in a tree like structure. Lower level information specifies upper level concept more closely. Under main metadata element there may be one or more sub-elements that may be split into one or more instance of information.

The main element, that is also the most top element in the final information model, is a *Process* without specifying the objective of it. Idea is that same information model may be used, with minor modifications, in other business areas too and not just for statistical purposes.

Process element is then further divided by its purpose, so as to create own branch for a *Statistical Process* or other types of processes. This node is the main point for the information model that covers both *statistical data* and process *phases*. In the figure 16, below, the main metadata elements are presented. These elements may also be called as concepts but for the clarity of text, the term “element” is more suitable in this context when the aim is to describe hierarchical structure of information.

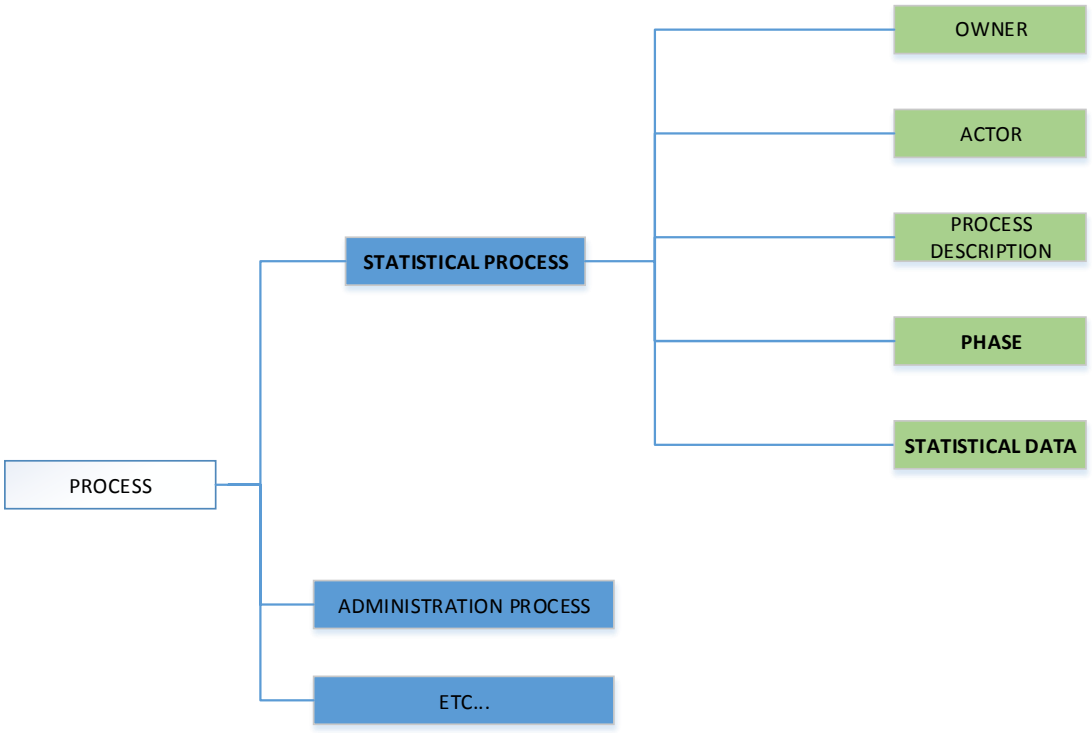


Figure 16. The main metadata elements in the final information model for knowledge-based system

The figure 16, shows how *Statistical Process*-element is further divided into five sub-elements; *Owner, Actor, Phase, Process Description* and *Statistical Data*, of which, from

the statistician point of view, most important information compositions are *Phase* and *Statistical Data*.

The other three sub-elements are needed to identify and to separate statistical processes from each other. Owner of the statistical process is commonly national statistical institution and named chief director in it, so this need to be stored in information model. Similarly it is important to define the actor of process and to know the person in charge. *Process Descriptions* include verbal and pictorial descriptions of process in general. The aim here is to provide enough information about the process without being too specific, at this point, in the details.

In the annex 5, is a complemented list of the main elements, their sub-elements and information fragments that are designed to the final information model.

Next two chapters handle the structure of *Phase*- and *Statistical Data* metadata elements more elaborately. It need to be noted that the UNECEs' GSIM metadata practices and instructions have been taken into account in the design of the final information model. The objective was to link the GSIM with the designed information model as well as possible. Linking is not thorough for some important process information is lacking from the international models.

6.2.2 The final information model for describing statistical data

Statistics Finland has common information model, the COSSI that is used mainly for describing statistical data (see the chapter 3.2.4.2 and the annex 4). It is also used as a common publication format and provides a format for survey questionnaire.

Hence, there is no need to design totally new metadata content instead validate content of the COSSI-model and only add new information elements in it when necessary. Validation of the COSSI-model revealed that it is mainly sufficient for describing statistical data. Few additional information fragments were needed to complete content of the data descriptions while totally new metadata element, *Physical_datadescription,* is needed for describing name, location, format and properties of actual datasets.

The figure 17 presents both of these sub-elements, *Data_description* and *Physical_datadescription,* that are needed to sufficiently describe content and location of statistical data.

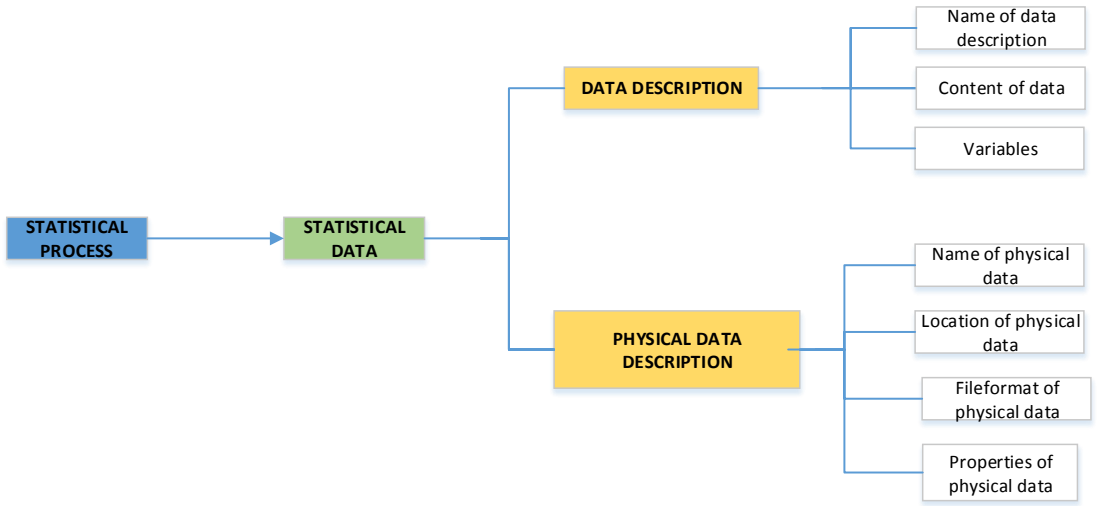


Figure 17. *Statistical Data* -metadata element, and its' top-level information structure

Another addition was made to the COSSI-model. According to the current state analysis it was recognized that statistician want more information about variables and their properties. *Variables* –element was expanded with *variable properties* that represent intended *variable type, length* and *allowed values*. Other added variable properties are: chance to flag primary variable and statistical unit-variable as well as separation of classification variables from research variables. In the figure 18 is presented the structure of the designed variable properties.

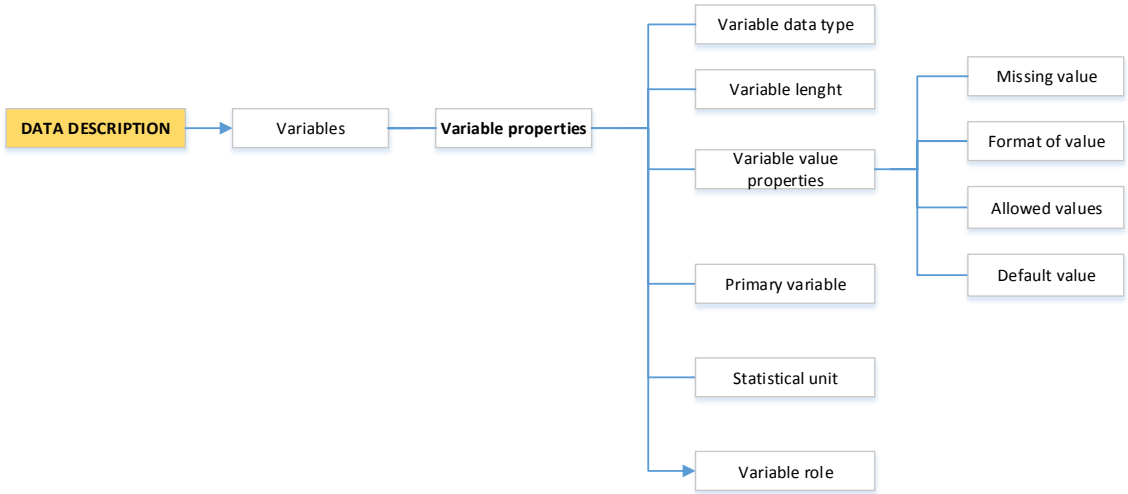


Figure 18. The structure of variable properties

Thorough list of the metadata elements, the sub-elements, information and concepts, that describe statistical data, is in the annex 5.

6.2.3 The final information model for describing statistics process

Design of an information model that describe sufficiently process information, its' phases and rules was more challenging. There were no previous model that could have been a starting point in the design. So, the design is merely based on the current state analysis-results, known best-practices, international recommendations and own interpretation.

While designing the sub-elements, it was obvious that each metadata element need to have an identifier and a purpose for its existence. Therefore each level of information has equivalently same sub-elements: *identification* and *description*. An example of this is demonstrated in the figure 19 that presents the sub-elements belonging to the metadata element *Phase*.

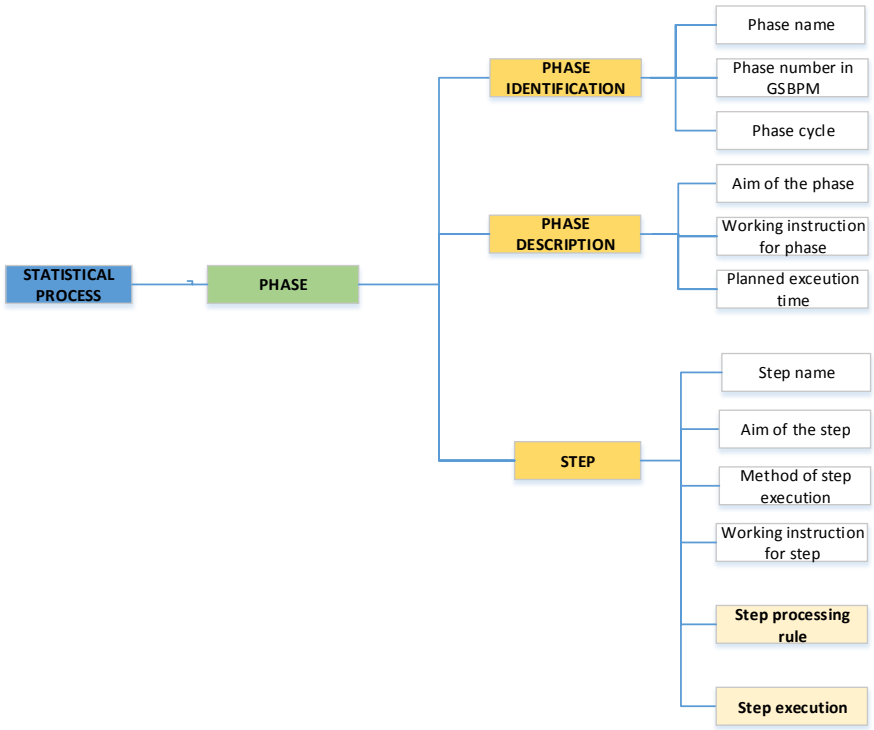


Figure 19. *Phase* metadata element and its' top-level information structure

The figure 19 shows how top-level metadata element, *Statistical Process,* is further divided into sub-elements. In this figure we concentrate only in *Phase*- element that is

needed for describing aim of phase, for describing methods that are used in phase execution and for describing individual steps in one specific phase.

At first, we need identify phase, so specific sub-element is defined for this purpose; *Phase identification.* Next, we need descriptive information that tells aim of specific phase, planned execution time and working instructions that are needed to execute task. This information is captured into sub-element *Phase description.* We also need information that specifies each step in one specific phase. This information is recorded in sub-element *Step.*

As said earlier, hierarchical structure accepts several instances of main- and sub-elements. This means that designed information structure allows statistician to record all process phases and their steps similarly.

7 Conclusions

This chapter summarizes the thesis process and outlines the purpose of the work. The chapter also describes how final step, from the initial proposal was taken in order to create the final information model.

7.1 The aim and the outcome of the research

Personal experience, in the field of statistics, has shown that very often there is too little standardised and structured information available for getting an overview to statistics production, to methods that are utilised in data processing and to quality of data and results.

Therefore personal ambition was to create an ideal information model that provides foundation for information management, structure for metadata management and outlines most important information that ought to be stored or captured from data processing. Idea was that this improved information covers at least those process phases where actual data is manipulated: Data collection, Processing, Analysing and Disseminating.

With this information model, a statistics production process may be described as a whole. Stored information is useful for statistician, methodologists, quality managers, researchers, analysts and other users who want to get an answer to question: *How data is treated in process and what is quality of results?*.

So, the objective of this research was to design an improved information model that is useful for describing different parts of statistics production: data, process, methods, rules, users, programming codes and other important issues.

The outcome, or a product, of this research is an improved information model that meets the set requirements. The improved information model notices the CSA-analysis results, the current best practices and the international development work in this field as well as the normative requirements and the international information models.

It follows standardised, hierarchical structure in metadata management and shows objects that are important for describing process and its' results. Outcome of this research is

> *An information model that takes totally new approach in describing information. Approach here is the process –oriented, not the data oriented as it is in other information standards.*

This model may be used for extending information content in the Statistics Finland's common metadata system and also for improving the current international standards.

Benefits of the final information model are

- It provides unified structure for describing statistical data and process related information in systematic way,
- It links tightly a process model, the GSBPM, with metadata,
- It takes advantage of the COSSI-model and expands it in those parts where additional information is necessary
- It follows, hierarchical information structure that is flexible enabling enlargement of structure and addition of new information fragments, according to the set requirements
- When in use and statistical information is stored based on the final information model, it provides key information for the national and international quality reporting

7.2 The description of how the initial model proposals are taken into account in final information model

This research process followed iterative approach that enabled improvement of the initial proposal further, towards the most suitable solution. Especially the feedback that was received from the focus group and thesis mentor, was very valuable even though it forced to reconsider the proposed model thoroughly once again.

It is obvious that the initial proposal and the final improved information model differs little when comparing at the structure of planned information. Both models include metadata elements that are divided into information fragments. By contrast, the final model refines the information structure further providing view also to the information that relates to statistical process. The final model also provides detailed descriptions for the planned concepts of the model.

Next, we take a look at the specific metadata elements that were presented in the initial proposal, but are slightly differently accomplished in the final model. The following list shows all metadata element that were designed for the initial proposal, and offer explanation of how these were solved in the final model.

- *DataCollection* – Data specific information and indicators are stored in the final information model as *Data Descriptions*, while process specific information will be stored in *Phase* descriptions.
- *Classifications* – These are kept unchanged because information content in the current Classification system satisfies current need.
- *Data Descriptions* – Data specific information is expanded with variable-specific properties and indicators. Beside data descriptions, a new metadata element called *Physical Data Description,* is defined so that metadata concerning actual dataset may be stored.
- *VariableQualityMeta, DataQualityMeta & Observation Meta* —Verbal descriptions of quality indicators are stored to the *Data Descriptions* while actual indicator values are stored in physical datasets.
- *DataContent*-- Data content specific information is stored in *Data Descriptions*
- *File Catalogue* – This information is not stored because it is easy enough to compose a list of files from existing *Data Descriptions* and *Physical Data Descriptions*.

- *PublishedTableDescription, PublishedFigureDescription & PublicationDescription -- PublicationDescription* is kept as it is in the current COSSI-model. *Data Descriptions* may be used for describing the tables and figures.
- *StatMethodMeta*—Introduced methods are specified in the metadata element: *Phase>>Step >> Step processing rules* –description.
- *ProcessingMeta* –This metadata element has very thorough descriptions in final model in metadata element called as *Statistical Process >>Phase*.
- *MonitoringMeta* -- This information is not stored separately because it is easy enough to compose monitoring information from existing descriptions and log-files[25].
- *WorkingInstructions* – These are stored to the metadata element Statistical *Process >> Phase >> Phase Descriptions* or *Step* depending on accuracy of instruction.
- *Methodological library* – This is suggested in the initial proposal and also in the final model
- *Operational Planning and Guidance System STOJ* – This is existing information that may be used as controlling information in process planning and monitoring
- *IT-architecture* – A generic IT-architecture is designed as instructed by the Government ICT Center[26], so information content describing IT-systems does not belong to this research
- *Background information (legislation, manuals, recommendations)*—All background material should be described in higher level working instructions in the metadata element *Process >> Statistical Process >> Process Description >>Process working instructions*

[25] Log-file is a text file that express what and how certain actions are executed on a computer, a server, a website etc. https://dictionary.cambridge.org/dictionary/english/log-file .
[26] The government ICT Center is a service center that provides ICT services for central government. http://www.valtori.fi/en-US

7.3 Ideas for developing and improving of current practices

As a result to this research, content of the final information model is planned and hierarchy of information is structured.

Beside the planned model, there are items that need further improvement or need to be created in order to successfully implement the planned information model in to the statistics production. These items are following:

1. *The common metadata system*
   1.1. the COSSI model need to be expanded with suggested metadata elements and information fragments
   1.2. all statistical methods need to be stored in a new methodology bank that includes existing concept library
2. *Mode of operations and common*, *system independent, processes* are needed for:
   2.1. extracting the metadata from the common metadata system and for updating metadata back in to the system,
   2.2. calculating quality indicators based on an information that is stored as a processing rules in the metadata system,
   2.3. checking data quality with generic analysis-processes utilising the stored metadata and executing an analysis in selected process steps,
   2.4. reporting process-, statistical data- and results quality, based on the stored metadata, indicator-values and monitoring information,
   2.5. converting information, such as the data descriptions and data values, in standardised format that may be delivered to the external users
3. *Standardisation of*
   3.1. A data storage systems like databases
   3.2. A in-house designed software and applications

The list above proves that several improvements ought to be done in order to meet the generic requirements

> to ensure the availability of reliable statistical information (Statistics Act 2004), and
>
> to base production of statistics on common and standardised processes and transforming raw data into statistical products according to generic and commonly accepted information concepts. (UNECE 2011b, p. 3-4)

7.4 Lessons learned

This research has shown how difficult it is to understand an information in its' pure form. Very often in discussions the meaning of information is mixed up with other issues and views. For example in the assessment of the initial proposal the participants brought forward ideas of how information may be stored to the database or utilised in process, instead of concentrating to the content of information that is needed for describing statistical process and data.

This example shows how challenging the topic is and how difficult it is to identify the essence of information. Only, when we identify information fragments that are needed in the detailed descriptions, we may model the structure of the statistical information. After this, it is time to plan how information is stored to the common systems, utilised in the process and monitored for quality. Not forgetting the common practices that are needed to demonstrate where and how stored information may be utilised.

References

References are listed according to their appearance:

Arofan, G. (2011), *The Data Documentation Initiative (DDI): An Introduction for National Statistical Institutes*. [Online] Available from: http://www1.Unece.org/stat/platform/display/metis/Existing+resources+related+to+the+relationship+between+SDMX+and+DDI [Accessed 5th August 2016].

Business Dictionary (n.d), *Information-- Definition*. [Online] Available from: http://www.businessdictionary.com/definition/information.html [Accessed 2nd April 2018].

Coghlan, D. & Brannick, T. (2014). *Doing Action Research in your own organization*. London:Sage Publications Ltd.

DDI Alliance (2016a), *DDI -- Document, Discover and Interoperate*. [Online] Available from: http://www.ddialliance.org/ [Accessed 10th February 2018].

DDI Alliance (2016b), *DDI -- Explore Documentation.* [Online] Available from: http://www.ddialliance.org/explore-documentation [Accessed 10th February 2018].

European Union (n.d.), *EU institution and bodies*. [Online] Available from: http://europa.eu/european-union/about-eu/institutions-bodies_en [Accessed 8th January 2017].

European Union (2010), *Legal framework for European statistics - The Statistical Law*. [Online] Available from: http://ec.europa.eu/eurostat/web/products-statistical-books/-/KS-31-09-254 [Accessed 2nd April 2018].

Eurostat (2013), *Glossary:Consumer price index (CPI)*. [Online] Available from: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:Consumer_price_index_(CPI) [Accessed 2nd April 2018].

Eurostat (2016 a), *Code of Practice*. [Online] Available from: 28th September 2011 http://ec.europa.eu/eurostat/web/products-manuals-and-guidelines/-/KS-32-11-955 [Accessed 6th February 2016].

Eurostat (2017), *Glossary:European system of national and regional accounts (ESA 2010)*. [Online] Available from: http://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:European_system_of_national_and_regional_accounts_(ESA_2010) [Accessed 2nd April 2018].

High, J. A. & Miller, B. J. (2007), *Projects, Programs and Links – Oh My!*. In:SAS Global Forum 2007. Paper 145-2007.

ISI (2009), *Declaration on Professional Ethics*. [Online] Available from: http://isi.cbs.nl/ethics0index.htm [Accessed 13th February 2016].

Johnson, G., Whittington, R., Scholes, K., Angwin, D. & Regnér, P. (2015). *Fundamentals of Strategy*. 3rd edn. Harlow:Pearson Educated Ltd.

Official Statistics of Finland (2013), *Suomen virallisen tilaston laatulupaus (Quality declaration of Official Statistics of Finland)*. [Online] Available from: http://www.stat.fi/meta/svt/svt_laatulupaus_2013.pdf [Accessed 2nd April 2018].

Praženka, D. & Boško, P. (2011), *Combining technical standards for statistical business processes from end-to-end.* [Online] Available from: https://ec.europa.eu/eurostat/cros/system/files/S4P4.pdf [Accessed 8th January 2017].

Public recommencation (2009), *JHS 171 ICT-palvelujen kehittäminen: Kehittämiskohteiden tunnistaminen (JHS 171 Improvement of ICT services: Identification of development areas)*. [Online] Available from: http://docs.jhs-suositukset.fi/jhs-suositukset/JHS171/JHS171.html [Accessed 8th January 2017].

Regulation (EC) No 223/2009 of the European Parliament and of the Council on European statistics. [Online] Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:087:0164:0173:en:PDF [Accessed 2nd April 2018].

Sdmx (2016), *SDMX Glossary*. [Online] Available from: https://sdmx.org/?sdmx_news=new-sdmx-glossary-available [Accessed 2nd April 2018].

Statistics Act 280/2004. [Online] Available from: http://www.tilastokeskus.fi/meta/lait/2013_tilastolaki_en.pdf [Accessed 2nd April 2018].

Statistics Finland (2003), *Common Structure of Statistical Information (CoSSI), Definition Description*. [Online] Available from: http://www.stat.fi/org/tut/dthemes/drafts/cossi_definition_descriptions_v_09_2003.pdf [Accessed 9th December 2016].

Statistics Finland/Rouhuvirta, H. (2010), *The Common Structure of Information Model*. Presentation: 9th April 2010.

Statistics Finland/Ollila, P. (2010a), *Kysely editointi ja imputointikäytänteistä (Questionnaire about the practices used for editing and imputing)*.
Presentation: 19th March 2010.

Statistics Finland/Ollila, P. (2010b), *KYSELY TILASTOKESKUKSEN TILASTOJEN EDITOINTI- JA IMPUTOINTIKÄYTÄNNÖISTÄ (Report on editing and imputing questionnaire-results).*

Statistics Finland/Ollila, P. (2012a), *Main phases in the Generic Editing Model*.
Presentation: 5th January 2012.

Statistics Finland/Ollila, P. (2012b), *Raaka-aineiston, editoinnin ja imputoinnin indikaattorit (luonnos)(Indicators for raw data, editing and imputing; draft)*.

Statistics Finland (2016a), *Resources*. [Online] Available from: http://tilastokeskus.fi/org/tilastokeskus/voimavarat_en.html
[Accessed 9th December 2016].

Statistics Finland Intranet (2016b), *Laadunhallinta ja ammattietiikka (Quality Management and Professional Ethics)*. [Online] Available from: http://intranet.stat.fi/tilastot-ja-tuotteet/laadunhallinta-ja-ammattietiikka/Sivut/default.aspx
[Accessed 8th January 2017].

Statistics Finland Intranet (2016c), *Tilastoprosessi (Statististics Process Model)*. [Online] Available from: http://intranet.stat.fi/tilastot-ja-tuotteet/tilastoprosessi/Sivut/default.aspx [Accessed 9th December 2016].

Unece (2010), *Exploring the relationship between DDI, SDMX and the Generic Statistical Business Process Model*. [Online] Available from: http://www.ddialliance.org/sites/default/files/ExploringRelationshipBetweenDDI-SDMX-GSBPM.pdf [Accessed 10th February 2017].

Unece (2011a), *Strategic Vision.* [Online] Available from: https://statswiki.Unece.org/display/hlgbas/Strategic+vision+of+the+HLG-MOS [Accessed 2nd April 2018].

Unece (2011b), *Business case for applying DDI and SDMX*. [Online] Available from: https://statswiki.unece.org/pages/viewpage.action?pageId=63930579 [Accessed 5th August 2016].

Unece (2012), *Toward GSIM V1.0 as a cornerstone for common reference architecture*. [Online] Available from: 6th November 2011 https://www.Unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.50/2012/02_Australia.pdf [Accessed 10th February 2018.

Unece (2013a), *Generic Statistical Business Process Model: GSBPM v5.0*. [Online] Available from: https://statswiki.unece.org/display/GSBPM/GSBPM+v5.0 [Accessed 8th January 2017].

Unece (2013b), *Generic Statistical Information Model (GSIM): Communication Paper for a General Statistical Audience*. [Online] Available from: http://www1.Unece.org/stat/platform/display/gsim/GSIM+Communication+Paper [Accessed 9th December 2016].

Unece (2013c), *The Generic Statistical Business Process Model*. [Online] Available from: 23rd December 2013 http://www1.Unece.org/stat/platform/display/metis/The+Generic+Statistical+Business+Process+Model [Accessed 2nd March 2016].

Unece (2013d), *Generic Statistical Information Model (GSIM): Implementing GSIM (Version 1.1, December 2013)*. [Online] Available from: 24th December 2013 http://www1.Unece.org/stat/platform/display/metis/Generic+Statistical+Information+Model [Accessed 10th March 2016].

Unece (2013e), *GSIM and standards*. [Online] Available from: 24th December 2013 https://statswiki.Unece.org/display/gsim/GSIM+and+standards [Accessed 10th March 2016].

Unece (2015), *Quality Indicators for the Generic Statistical Business Process Model GSBPM) Version 5.0*. [Online] Available from: 7th December 2017 https://statswiki.unece.org/display/GSBPM/Quality+Indicators+Home [Accessed 2nd April 2018].

United Nations Statistics Division (2014), *Fundamental Principles of National Official Statistics*. [Online] Available from: 29th January 2014 http://unstats.un.org/unsd/dnss/gp/fundprinciples.aspx [Accessed 13th February 2016].

Web Crawler (2016). *Wikipedia* [Online]. Available from: https://en.wikipedia.org/wiki/Web_crawler [Accessed 8th January 2017].

APPENDICIES

The CSA -questionnaire

Background questions

All replies are handled confidentially. The only identification information stored are classification variables. No other identification information is saved. At first few questions about the respondents.

1. In which statistics department are you working? Only one option may be selected.
    - TY Economic and Environmental Statistics
    - YY Business Statistics
    - VE Population and Social Statistics
    - Other
2. Which kind of statistics are you working with? One or several options may be selected.
    - Monthly statistics
    - Quarterly statistics
    - Annual statistics
    - If other; please describe it
3. If you want to join the free cinema ticket lottery, please give your name?

Questions about statistics production process.

Detailed information about generic process model is in intranet. (*link to the page)*

4. Do you know the Generic Statistical Business Process Model (GSBPM)?
    - Yes
    - No
5. Have you at your work done process descriptions from one process phase or whole process according to the GSBPM?
    - Yes
    - No
6. Which process phases have you described according to the GSBPM?
    - Collection
    - Process
    - Analyse
    - Publication

7. Have you described, for example with Ms Excel, metadata objects used in statistics production process?
Here metadata refers to parameters, formulas, filtering criterias.

- Yes
- No

8. Could you give a sample of essential metadata objects used in your statistics production?
You are encouraged to send list of metadata objects by email to kristiina.nieminen@stat.fi

Questions about data descriptions.

Detailed information about data description practices in Statistics Finland is in intranet. (*link to the page)*

9. Have you at your work created data descriptions (information about data and variables in it) with Ms Excel or with MuuttujaEditori[27]?

- Yes
- No

10. Have you at your work created classifications with Ms Excel or updated existing classifications with LuokitusEditori[28]?

- Yes
- No

11. Do you recognise process phases in statistics production where these descriptions may be exploited?

- Yes
- No

[27] MuuttujaEditori is in-house software used for describing statistical information in a standard manner and uniform practices. Data definitions are stored in XML-format to central metadata database.
[28] LuokitusEditori is in-house browser application used for describing classifications used in statistics production. It provides functionality to record and observe existing classification.

12. In which process phase your statistics is using data descriptions?

- Collection
- Process
- Analyse
- Publication
- Do not know
- Other; what?

13. In which phase are you using classifications stored in common database or somewhere else?

- Collection
- Process
- Analyse
- Publication
- Do not know
- Other; what?

14. Do you know the content of COSSI-model structure?

- Yes
- No

Open questions

Finally I will ask you to answer to some open-text box question concerning statistical information (processmeta, quality indicator, data descriptions etc.) already produced along the process or you feel need to get out of process.

15. What kind of metadata, like quality indicators, process passing times, response rates etc, you are already producing from the statistical data or of the process?

16. What kind of metadata you think should be collected from process phases in order to offer enough information for statistician about the data quality and the process? Give 2-5 example for each process phase.

- Collection
- Process
- Analyse

- Publication

17. What kind of status reports or quality reports you would like to get from statistics production process by process phase?

- Collection
- Process
- Analyze
- Publication

The CSA conclusions

Results from the web-questionnaire-analysis

***Conclusion 1**:* The table 9 prove that implementation of the generic process model GSBPM is underway in Statistics Finland although this not yet comprehensively adapted. Only six respondents, approximately 15 % of all, has described all process phases according to the GSBPM.

***Conclusion 2**:* There are differing ways to store metadata at statistical units. Some define metadata with Ms Excel while others define the important metadata in text documents or in the programming code. Therefore the uniform practices and structured information model and -system is needed to enable statistician to record metadata, to observe metadata content and to use stored metadata in their process

***Conclusion 3**:* The figures in table 10 confirm assumption that the data descriptions and classifications are well-known and common practices are used at statistical departments.

***Conclusion 4**:* The main findings, so far, may be summarised shortly: The awareness of the GSBPM, the data descriptions and the classifications is good. Yet common practices need to be developed and trained in order to implement this standardised approach throughout the statistical departments and unit.

***Conclusion 5**:* The awareness of the COSSI-model is moderate so more training is needed in order to ensure that statistician clearly understand the utilities of the COSSI model. A comprehensive understanding is needed especially if statistician works as a process developer or a senior adviser. More you understand the content of common information models and standards, more easily you identify where this information may be captured and utilised.

***Conclusions 6**:* The results show that the statistician are willing to share their current practices (question 16) and they know very well their requirements for more detailed information (question 17). This confirm the assumption that more information is needed from the statistical data processing about the processed data, the introduced statistical

methods, the decision rules, the boundary values and most of all about the quality of the results. The total list of replies is very comprehensive offering many ideas for the building of the initial proposal.

Results from the case 1. The results of Administered Data Collection-analysis

Conclusion 7*:* The described practices are very useful, yet metadata enlargement is quite limited. Statistician need more information from an order and a supplier. Details describing the order are an identifier of order agreement, location of an order, date of an agreement and content of an agreement. Information about the supplier are contact details like a name of an organisation, a department responsible for an agreement, a department responsible for a data-transmission and contact person details.

Conclusion 8*:* Practices that are used for data analysis are useful, yet slight improvements need to be done to achieve proper functionality. At the moment calculations are accomplished for all variables of data. This approach consumes resources too much. So a primary variable of statistical data need to be defined to the data description in order to use this information for limiting calculations.

Case 2. The Generic Editing Model

Conclusion 9*:* All these indicators need to be taken into account in the initial proposal design. Measures describing and refining data are essential for the quality reports. The web-questionnaire, that was organised in 2010 (Ollila, P., 2010a), provided a view to the information that the methodologists value.

Conclusion 10*:* A methodology bank that contains information about the statistical methods and concepts need to be designed and implemented in Statistics Finland. The methodology bank, if founded and updated by official statistical office, is useful for the internal users but also for external users, like analysts, students and a teachers.

Results from the case 3. The big data processing-analysis

Conclusion 11*:* At the moment IT-related and content related information is not yet stored in unified and structured form, instead statisticians store this kind of information with a varying methods. Therefore unified method, to store and to utilise the process

and data related information, is needed. Current content in the common metadata system is too limited, so it need to be enhanced in order to let statistician record also IT-related information systematically.

Results of Systematic Quality Audit–report analysis

Conclusion 12*:* These results again confirm the need for more detailed information. Same perspectives, as was seen in the questionnaire analysis findings are presented here.

The Questionnaire for the assessment of the initial proposal

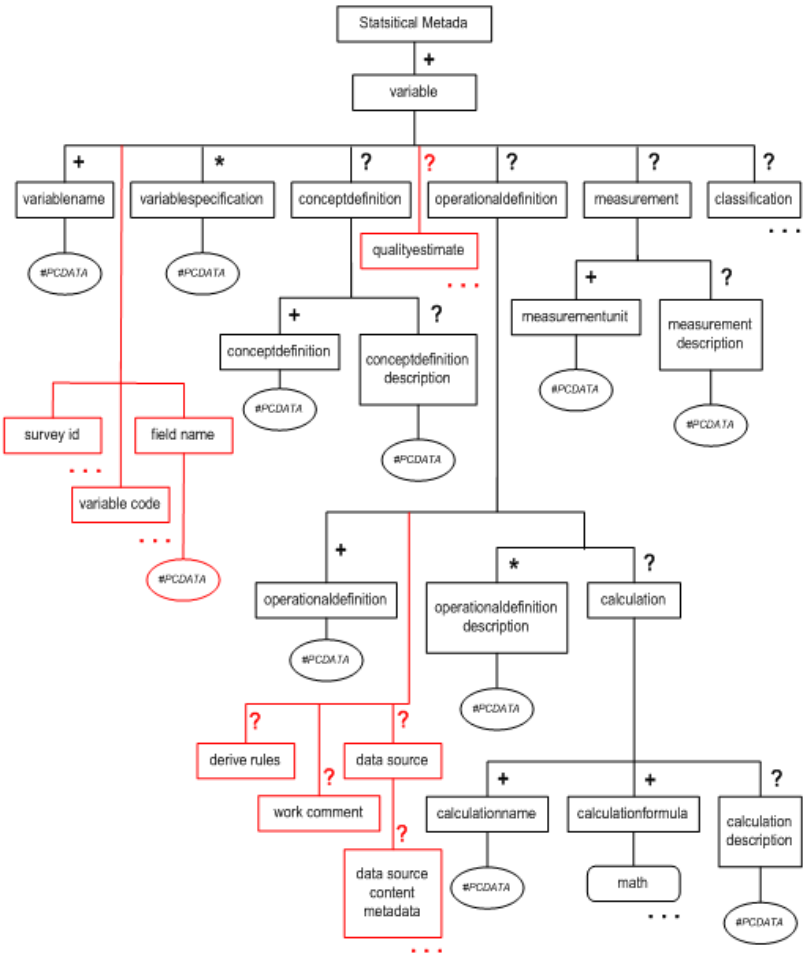
17.3.2017 FEEDBACK FOR THE INITIAL PROPOSAL – GROUP 1

Background information of respondent	Department / Unit:
	Working years in Statistics Finland:
	Sex/Age:

Please circle the most suitable grade (1-10) that correspond with your opinion. Replies to the open text questions are given to the field on the right side of question.

Questions	**Grading** Unsuitable =1									Good=10
1. Is the background job done with sufficient scope?	1	2	3	4	5	6	7	8	9	10
2. Which aspects should have been taken in the current state analysis?										
3. Has researcher orientated to the source material enough?	1	2	3	4	5	6	7	8	9	10
4. Which additional source material should researcher read through										
5. What grade do you give to the initial proposal?	1	2	3	4	5	6	7	8	9	10
6. What good is in the initial model?										
7. What defects does the initial model include?										
8. Which aspect is missing completely?										
9. Which aspect has got too much emphasis?										
10. Other feedback …										

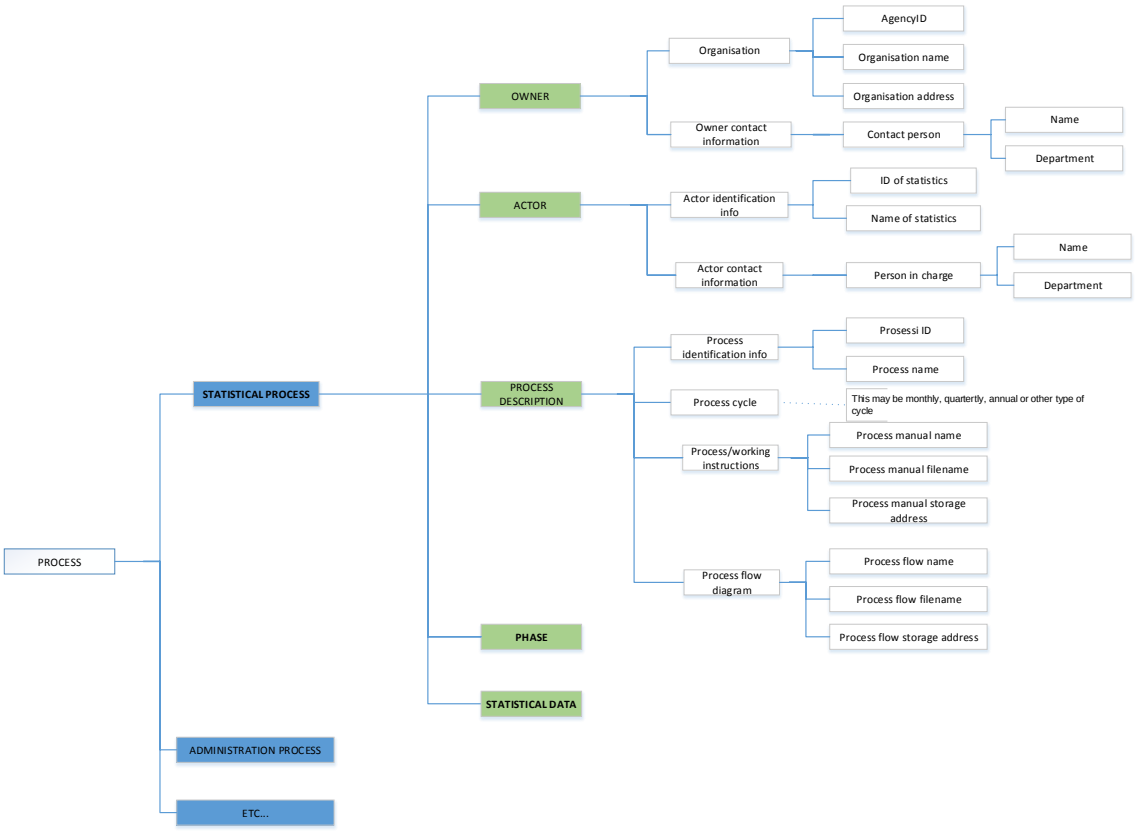
The COSSI, logical concept model-structure



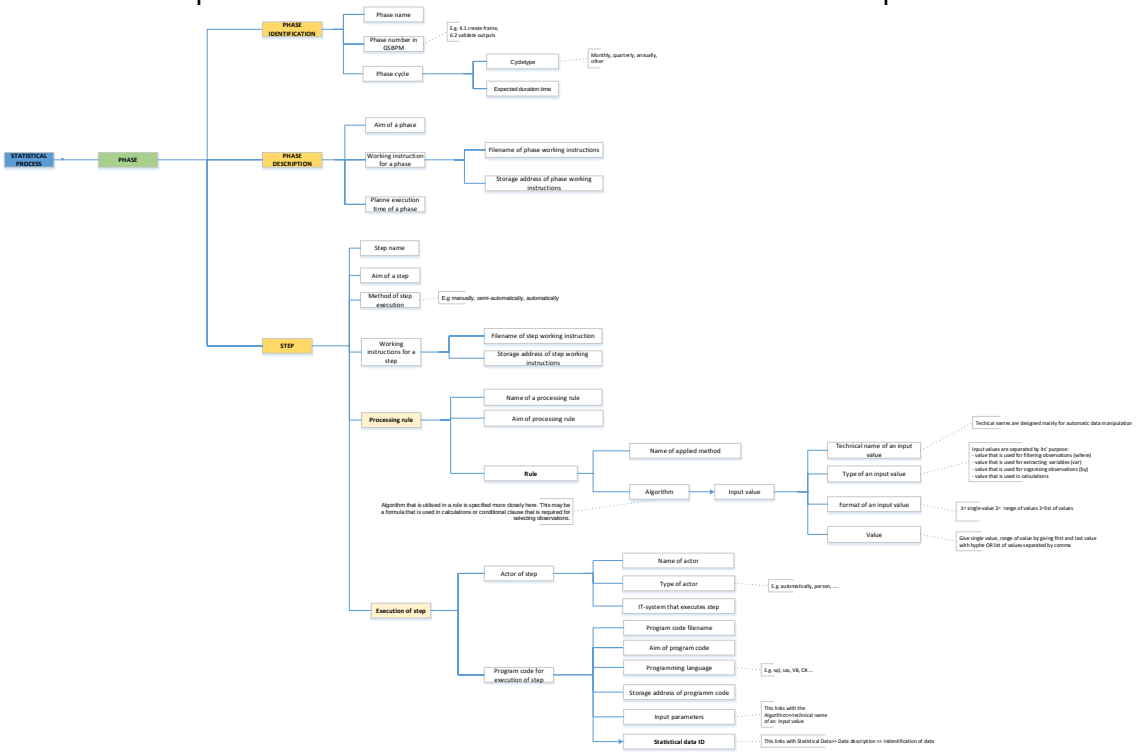
Source: Statistics Finland (2010)

The Final information model for knowledge-based work

Main metadata elements and their sub-elements

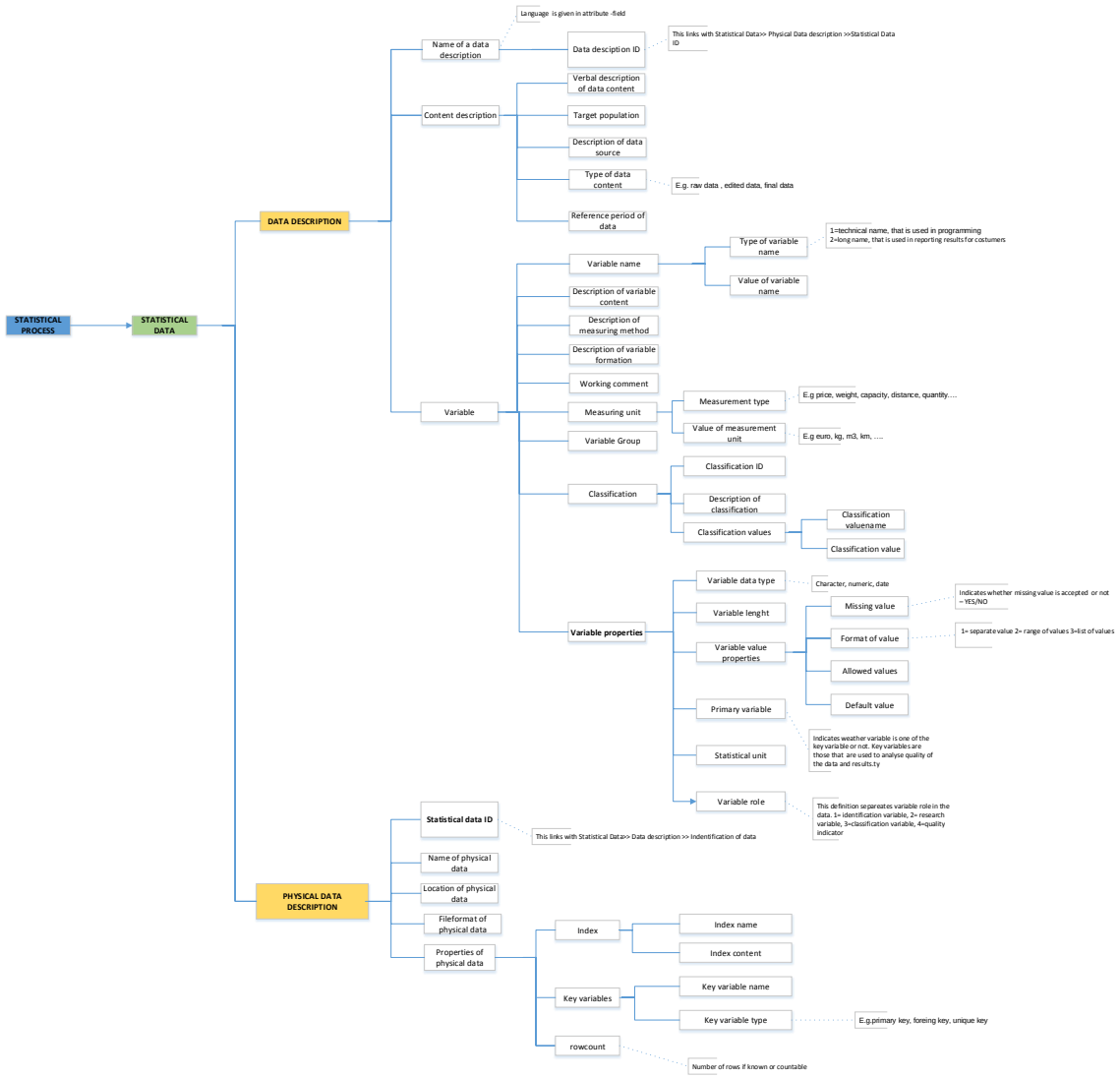


Detailed description of main metadata element – PHASE and concept definitions



hierarchy	Concept/metadata element	Description of metadata element. If there is already a description for an element, then it is listed in colums, GSIM and OECD, on the right side.	GSIM	OECD
	Process	Process is a set of repeated or on-off actions that are performed in order to transfom inputs into outputs.		
0	**Statistical Process**	Statistical process is an array of actions, repeated or non-recurring, that are perfomed in order to produce results in a form of statistics from input information . Typically statistical process cover process phases such as data collection, processing, analysis and dissemination of results. Statistical procss is further divided into Owner, Actor, Process Description and Phase. -	Statistical program cycle	Statistical Processing : The processes for manipulating or classifying statistical data into various categories with the object of producing statistics. (OECD)
1	**Owner**	Process owner is a national statistical institution and a named person who is responsible of statistical processes, results of processes and resources used to accomplish the aim, statistics. Owner is further divided into Organisation and Owner contact information.	Agent	
1	**Actor**	Process actor is an statistical unit and named statistician that is responsible for the planning of a process, implementation of process steps and tasks, and for ensuring the results. Actor is further divided into Actor identification and Actor contact information.		
1	**Process Description**	Process description is used for outlining the purpose and the aim of process and to give general view to the needed process phases. Process flow is presented graphically and complemented with working instructions. Process Description is further divided into Process identification info, Process cycle, Process working instructions and Process flow diagram.		
1	**Phase**	Statistical process is divided according to the GSBPM into eight phases that are all described, one by one, by their main features. This metadata element is one of the key elements that is further divided into lower level descriptions that specify higher level description. Phase is further divided into Phase Identification, Phase description and Step.	Business Process: he set of Process Steps to perform one of more Business Functions to deliver a Statistical Program Cycle or Statistical Support Program.(GSIM)	
2	**Phase Indentification**	Each process phase of a statistical process has identification details in order to separate it from the other phases. This element also links phase with phases in the GSBPM model, specifies cycle of phase (e.g. annual, quarterly, monthly) and expected duration of phase. Phase Identification is further divided into Phase name, Phase number in GSBPM and Phase cycle.		
2	**Phase Description**	Phase descriptions are used complement upper level process description. The aim is to give more detailed information of phase, such as the aim of phase, working instructions and planned execution time of phase. Phase Description is further divided into Aim of a phase, Working instruction for a phase and Planned execution time of a phase.		
2	**Step**	Step is an action that is performed in order to carry out task dedicated for it. Each phase is further divided into steps that are more specifically described, one by one, under Phase-element. Detailed, step-specific working instructions and processing rules are described in this element that is further divided in to sub-elements. Step is furter divided into information that specify step, Processing rule and Execution of step.	Process Step:Process Steps can contain "sub-steps", those "sub-steps" can contain further "sub-steps" within them and so on indefinitely. Typically, the outputs of one Process Step become inputs to the next Process Step. There can also be conditional flow logic applied to the sequence of Process Steps, based on parameters which have been passed in, or conditions met by the outputs of a previous Process Step.(GSIM)	Activity
3	**Processing rule**	Processing rule describes more closely how a certain step ought to be executed. This presentes information on the purpose of rule and applied methods. Processing Rule is further divided into Name of a rule, Aim of a rule and description of applied rule is defined in element Rule.	Process Method:A specification of the technique which will be used to perform the unit of work. (GSIM)	
4	**Rule**	Rule describes applied method and algorithm that is used in it. It also describes input values that are allowed or may be used in calculation or used for filtering and organising observations, and for subsetting variables. Rule is furter divided into Name of applied method and Algorithm.	Rule:A specific mathematical or logical expression which can be evaluated to determine specific behavior.(GSIM)	
5	**Algorithm**	Algorithm is a rule that is used in calculations of derived variable or used as a conditional clause that is required when selecting observations. Algorithm has only one sub-element that is Input Value.	Algorithm: The rule expressed as an algorithm (GSIM)	
6	**Input value**	Input value is defined for all those algorithms that are described. One algorithm may take as an input, in a form of a parameter, one or more input values. These stored input values are used for extracting the observations (where) or variables (var), for organising dataset (by) or to perform calculation. Input value is further divided into Technical name of an input value, Type of an input value, Format of an input value and Value (itself).		
3	**Execution of step**	Execution of step describes mainly actor of specific step and program code that is needed to perform step. In modern environment there is very few step or task that may be performed without a program code, so it need to be described here. This element is further divided into Actor and Program Code.		
4	**Actor**	Actor is a person or IT-system that performs planned step at the specified time as	Individual:A person who acts, or is designated to act towards a specific purpose.(GSIM)	
4	**Programming Code**	Programming code is used for accomplishing planned phase, task or step. It may be executed automatically or manually. The aim is to describe name and location of a code and give details about the nature of a code.		Computer Program:A set of instructions directing the computer which operations to perform. (OECD)

Detailed description of main metadata element – STATISTICAL DATA and concept definitions



hierarchy	Concept/metadata element	Description of metadata element. If there is already a description for an element, then it is listed in colums, GSIM and OECD, on the right side.	GSIM	OECD
	Process	Process is a set of repeated or on-off actions that are performed in order to transfom inputs into outputs.		
0	**Statistical process**	Statistical process is an array of actions, repeated or non-recurring, that are perfomed in order to produce results in a form of statistics from input information . Typically statistical process cover process phases such as data collection, processing, analysis and dissemination of results. Statistical procss is further divided into Owner, Actor, Process Description and Phase. -	Statistical program cycle	Statistical Processing : The processes for manipulating or classifying statistical data into various categories with the object of producing statistics. (OECD)
1	**Statistical data**	Statistical data is all those datases that are one way or other utilised in a statistical process. Dataset may contains survey observations or aggregated data: so it may be microdata or macrodata. There is no difference in describing microdata compared to macrodata. All datasets or even database tables are described in similar manner.		Basical statistical Data : Basic statistical data are data collected on a regular basis (by survey from respondents, or from administrative sources) by survey statisticians in the national statistical system to be edited, imputed, aggregated and/or used in the compilation and production of official statistics. (OECD)
2	**Data description**	Data description are needed for describing verbally data content. This information complements physical datasets where all recorded observations and their values may be observed. Data descriptions define more closely the content of dataset, lists all variables in it, specifies properties of variable, shows classifications that are linked with variable values. All variables are specified one by one. The aim is to provide enough information for statistician or researcher who aims to process data further in to statistics or survey results. Data description is further divided into identification information, description of DataContent and Variables.		
3	**Name of a data description**	Name of data description is defined in order to recognise and to separate data descriptions from each other. This sub-element has only one complementary information that is Data Description ID.	Input/Output-data: -- Any instance of an information object which is supplied to a Process Step Instance at the time its execution is initiated.Any instance of an information object which is produced by a Process Step as a result of its execution. (GSIM)	
3	**Content description**	Content description gives general view to the content of dataset. It answer to the questions such as what is target population, how is data collected, what is type of dataset (raw, edited, final) and what is reference period of dataset if this may be expressed. This element is further divided into Verbal description of data content, Target population, Description of datasource, Type of data content and Reference period of data.		
3	**Variable**	Variable -element describes each variable, one by one.Variable specific information cover: name and content of variable, how variable is measured in a survey, what is measuring unit applied to variable and what is classification applied to variable, etc. This element include several sub-elements, such as Measuring unit, Classification and Variable properties, that are divided further into more specific information fragments.	Variable	Concept
4	**Measuring unit**	Measuring unit describes the unit that is applied in measuring certain phenomenon that is expressed as variable. Measuring unit is further divided into Measurement type and value of measurement unit.		Unit of a measure:.A Unit of measure is the actual unit in which the associated values are measured (OECD)
5	**Classification**	Statistical classification: A Statistical Classification is a set of Categories which may be assigned to one or more variables registered in statistical surveys or administrative files, and used in the production and dissemination of statistics. Classification is further divided into Classification ID, Description of Classification and Classification values that is further divided into valuename, and code or value.	Statistical classification: A Statistical Classification is a set of Categories which may be assigned to one or more variables registered in statistical surveys or administrative files, and used in the production and dissemination of statistics. The Categories at each Level of the classification structure must be mutually exclusive and jointly exhaustive of all objects/units in the population of interest (GSIM)	
5	**Variable properties**	There is a lot of variable properties related information that ought to be stored. For example variable type, lenght and information that specifies how variable is treated in a process; e.g is variable very important for evaluating results (>>primary variable), or does variable include identification information for statistical unit (>>statitical unit) or is variable used as an research variable, classifcation variable or as quality indicator (>>variable role).		Statistical Unit: Statistical units are the entities for which information is sought and for which statistics are ultimately compiled. These units can, in turn, be divided into observation units and analytical units. The statistical units in the International Standard Industrial Classification (ISIC) Rev. 3 -- Indicator: A statistical indicator is a data element that represents statistical data for a specified time, place, and other characteristics.
2	**Physical Data Description**	Beside the verbal data description also physical dataset need to be described in order to know what is name and location of dataset, in which format dataset is stored in common storage system and what are specified properties of datasets. Physical Data Description is further divided into Name of physical dataset, Location of physical dataset, Fileformat of physical dataset and Properties of physical dataset. This element is further divided into identification information, name and location of a dataset, information that specifies properties of dataset.		Dataset is any organized collection of data (OECD).